AN EVALUATION OF STATISTICAL METHODS FOR DETERMINING AGREEMENT AND RELIABILITY IN MEDICINE

RAFDZAH AHMAD ZAKI

THESIS SUBMITTED IN FULFILMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PUBLIC HEALTH

FACULTY OF MEDICINE
UNIVERSITY OF MALAYA
MALAYSIA

2013

Original Literacy Work Declaration

Name of Candidate: RAFDZAH AHMAD ZAKI (780607-10-5854)

Matric No: MHC090010

Name of Degree: Doctor of Public Health (DrPH)

Title of Thesis: AN EVALUATION OF STATISTICAL METHODS FOR DETERMINING AGREEMENT AND RELIABILITY IN MEDICINE

Field of Study: Public Health (Epidemiology and Biostatistics)

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:

Designation:

Abstract

Background and objective:

Agreement and reliability are both important parameters in determining the quality of an instrument. The general aim of this study was to evaluate different statistical methods used to assess agreement and reliability of medical instruments that measure the same continuous outcome. This study compares the most commonly used statistical methods, and also compares the proposed method in the analysis of agreement. Two separate systematic reviews were performed at the beginning of this study to identify the most popular method used to assess agreement and reliability of medical instruments.

Methods:

Two systematic reviews, one on agreement studies and another on reliability studies were carried out. A cross-sectional study was then conducted to collect data in two population settings; an institutional and community setting. Data were collected for blood glucose level, systolic blood pressure, diastolic blood pressure, body weight, and peak expiratory flow rate to assess methods used in agreement studies. Data were also collected for reliability studies. The variables for this were systolic blood pressure, diastolic blood pressure, heart rate, body temperature, peak expiratory flow rate and carbon monoxide level. Evaluations of agreement and reliability statistical methods were carried out on the original clinical data and simulated data. The agreement evaluation involved a comparison of the most commonly used methods (Bland-Altman Limits of Agreement and Intra-class Correlation Coefficient for agreement), comparison of slopes and y-intercepts analysis, and a new proposed agreement model. The reliability evaluation involved a comparison of the Bland-Altman Limits of Agreement, Intra-class Correlation Coefficient for consistency (ICC_C), and Intra-class Correlation Coefficient for absolute agreement (ICC_A).

Results and Conclusion:

The systematic reviews identified some issues related to method comparison studies including the application of inappropriate statistical methods, and the importance of education in method comparison studies among medical professionals. The evaluations of different statistical methods provided different conclusions on agreement and reliability. Each method was found to have its own strengths and weaknesses and no single method was found to be perfect. Several recommendations were made including optimal sample size for Bland-Altman analysis, and a proposed flowchart to guide analysis of agreement and reliability in method comparison studies.

Abstrak

Latar Belakang dan Objektif Kajian:

Ketepatan (*agreement*) dan kepersisan (*reliability*) adalah dua parameter penting dalam menentukan kualiti alat-alat perubatan. Tujuan umum kajian ini adalah untuk menilai beberapa kaedah statistik berbeza yang telah digunakan untuk menilai ketepatan dan kepersisan alat-alat perubatan yang mengukur pembolehubah kuantitatif yang sama unit. Kajian ini membandingkan kaedah statistik yang paling banyak digunakan dibidang perubatan, dan juga membandingkan kaedah yang dicadangkan dalam kajian ini. Dua ulasan sistematik (*systematic review*) yang berasingan telah dilakukan pada awal kajian ini untuk mengenal pasti kaedah yang paling popular digunakan untuk menilai ketepatan dan kepersisan alat-alat perubatan.

Kaedah:

Dua ulasan sistematik (*systematic review*) telah dijalankan, satu berkenaan kaedah statsitik menilai ketepatan (*agreement*) dan satu lagi berkenaan kaedah statsitik menilai kepersisan (*reliability*). Satu kajian *cross-sectional* telah dijalankan untuk mengumpul data dari dua populasi (institusi dan masyarakat). Data yang telah dikumpulkan untuk menilai kaedah yang digunakan dalam kajian ketepatan ialah: paras glukosa darah, tekanan darah sistolik, tekanan darah diastolik, berat badan, dan *peak expiratory flow rate* (PEFR). Data yang dikumpulkan untuk kajian kepersisan ialah: tekanan darah sistolik, tekanan darah diastolik, kadar jantung, suhu badan, *peak expiratory flow rate* (PEFR) dan tahap karbon monoksida (CO). Penilaian kaedah statistik ketepatan dan kepersisan telah dijalankan pada data klinikal dan data simulasi. Penilaian ketepatan melibatkan perbandingan kaedah yang paling biasa digunakan (*Bland-Altman Limits of Agreement* dan *Intra-class Correlation Coefficient for absolute agreement*), *comparison of slopes and y-intercept analysis*, dan *agreement model* yang dicadangkan. Penilaian

kepersisan melibatkan perbandingan *Bland-Altman Limits of Agreement, Intra-class Correlation Coefficient for consistency* (ICC_C), dan *Intra-class Correlation Coefficient for absolute agreement* (ICC_A).

Hasil Kajian dan Kesimpulan:

Kedua-dua ulasan sistematik (*systematic review*) telah mengenal pasti beberapa isu yang berkaitan dengan kajian perbandingan kaedah (*method comparison studies*) termasuk penggunaan kaedah statistik yang tidak sesuai, dan kepentingan pendidikan dalam bidang kajian perbandingan kaedah (*method comparison studies*) di kalangan pengamal perubatan. Kaedah statistik yang berlainan memberikan kesimpulan yang berbeza dalam menetukan ketepatan dan kepersisan alat-alat perubatan. Setiap kaedah mempunyai kekuatan dan kelemahan sendiri, dan tiada satu kaedah yang sempurna. Beberapa cadangan telah dibuat termasuk sampel optimum bagi Bland-Altman analisis dan carta aliran yang dicadangkan untuk membantu penyelidik dan pengamal perubatan dalam analisis ketepatan (*agreement*) dan kepersisan (*reliability*) alat-alat perubatan dalam kajian perbandingan kaedah (*method comparison studies*).

Acknowledgements

I would like to express my utmost gratitude to my supervisor, Professor Dr Awang Bulgiba for the support, guidance and assistance throughout my research. My thanks also to my co-supervisor, Professor Dr Noor Azina Ismail, for her most helpful support and comments, especially the statistical guidance and advice. Her motivation and constructive comments during the many discussions were much appreciated. It was a great pleassure to conduct this research under both supervisors.

I would like to acknowledge with much appreciation the coordinator and staff of the UM Wellness Programme for providing facilities and assistance during my data collection that made this project possible. Special thanks to Dr Siti Munira whose help was invaluable during the period of my data collection. Special thanks also to the participants of this research i.e. from the University of Malaya Welllness Programme, the villagers of Kampung Teluk Gadong Kecil, Klang Selangor, and participants from the health screening program in Mid Valley Mall Kuala Lumpur.

Special acknowledgements to the University of Malaya for the grant that has enabled me to carry out this work (University of Malaya student research grant: PS162/2009B). Also thank you to the head and staff of Julius Centre University of Malaya (JCUM) who assissted in my work directly or indirectly, especially for the funding of my first journal publication under the University of Malaya/Ministry of Higher Education (UM/MOHE) High Impact Research Grant (Grant number E000010-20001).

I would like to give special thanks to both my parents for their constant support and encouragement. Last but certainly not least, special thank you to my husband Mr Faisal Tan who being supportive and understanding. This thesis is dedicated to my eldest son (Imran Tan, 7), my twins (Aisyah Tan and Aliyah Tan, 6), and especially to Sarah Tan (born 1st March 2011) and Fatima Tan (born 18th April 2013) for being part of me while completing this project and thesis.

Publications

The following papers have been published, submitted or presented from this thesis:

Journal:

1. Zaki, R., Bulgiba, A., Ismail, R., & Ismail, N. A. (2012). Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review. *PLoS ONE, 7: e37908.doi:10.1371/journal.pone.0037908.*
2. Zaki, R., Bulgiba, A., & Ismail, N. A. Testing the agreement of medical instruments: Overestimation of bias in the Bland-Altman analysis. Accepted for publication by Preventive Medicine.
3. Zaki, R., Bulgiba, A., Nordin, N & Ismail, N. A. A systematic review of statistical methods used to test for reliability of medical instruments measuring continuous variables. Accepted by Iranian Journal of Basic Medical Science.
4. Zaki, R., Bulgiba, A., & Ismail, N. A. Optimal Sample Size for the Bland-Altman analysis. Submitted to Journal of Clinical Epidemiology.

Conference/seminar:

1. Zaki, R., Bulgiba, A., Ismail, R., & Ismail, N. A Review of Statistical Methods to Assess Agreement in Medicine. Poster presentation: 31st Annual Conference of the International Society for Clinical Biostatistics, Montpellier, France. 29 August – 2 September 2010.
2. Zaki, R., Nordin, N, Bulgiba, A., & Ismail, N. Assessing Methods To Determine Reliability Of Medical Instruments. Oral presentation: 2nd International Conference On Quantitative Sciences And Its Applications (ICOQSIA2010), Penang. 2 – 4 April 2010.

3. Zaki, R., Bulgiba, A., Nordin, N., Ismail, R., & Ismail, N. Knowledge on Methods of Validation Study In Medicine. Oral Presentation: 42th Asia Pasific Academic Consortium for Public Health (APACPH) Conference, Bali, Indonesia. 24 – 27 November 2010.
4. Zaki, R., Bulgiba, A., & Ismail, N. A. Sample Size for Bland-Altman Analysis in Validation Study. Oral presentation: Asia-Link Clinical Epidemiology & Evidence Based Medicine (CEEBM) Conference, Bali, Indonesia. 27 – 28 November 2010.
5. Zaki, R., Bulgiba, A., & Ismail, N. A. Testing the agreement of medical instruments: Overestimation of bias in the Bland–Altman analysis. 1st Asia Pacific Clinical Epidemiology & Evidence Based Medicine (APCEEBM) Conference, Kuala Lumpur. 6 – 8 July 2012.
6. Zaki, R., Bulgiba, A., & Ismail, N. A. The Application of Intra-Class Correlation Coefficient (ICC) in Assessing the Reliability of Medical Instruments Measuring Continuous Outcomes. Accepted for poster presentation in the Faculty of Medicine Research Week, University of Malaya, Kuala Lumpur. 21 – 23 January 2013.

Table of Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

ABPM	:	Ambulatory Blood Pressure Monitoring
ANCOVA	:	Analysis of covariance
ANOVA	:	Analysis of variance
BP	:	Blood Pressure
bpm	:	beat per minute
BSA	:	body surface area
CI	:	Confidence Interval
CO	:	Carbon Monoxide
DBP	:	Diastolic Blood Pressure
DF	:	degree of freedom
EBM	:	Evidence-Based Medicine
ECG	:	Electrocardiogram
f^2	:	effect size
GCP	:	Good Clinical Practice
H_0	:	Null Hypothesis
H_1	:	Alternative Hypothesis
HBPM	:	Home Blood Pressure Monitoring
HR	:	Heart Rate
ICC	:	Intra-Class Correlation Coefficient
ICC_A	:	Intra-Class Correlation Coefficient for Agreement
ICC_C	:	Intra-Class Correlation Coefficient for Consistency
IPPP	:	Institute of Research Management and Monitoring
kg	:	Kilogram
l/min	:	litre per minute

LoA	:	Limits of Agreement
mmHg	:	millimetre of mercury
mmol/l	:	millimoles per litre
MS	:	Mean Square
MS_B	:	Between-Subjects Mean Square
MS_E	:	Error Mean Square
MS_S	:	Subject Mean Square
MS_T	:	Trials Mean Square
MS_W	:	Within-Subjects Mean Square
n	:	Numbers of sample in a certain condition
N	:	Total sample size
NICE	:	National Institute of Clinical Excellence
OLP	:	Ordinary Least Product
OLS	:	Ordinary Least Square
PEFR	:	Peak Expiratory Flow Rate
ppm	:	parts per million
PPP	:	Postgraduate Research Grant
p	:	*p*-value (probability of type I error)
r	:	Pearson correlation coefficient
r^2	:	coefficient of determination
$S_{\beta 1-\beta 2}$	:	standard error of the difference between slopes
SBP	:	Systolic Blood Pressure
SD	:	Standard Deviation
SE	:	Standard Error
SS	:	sum of squares
Temp	:	Body Temperature

UM	:	University of Malaya
UMMC	:	University Malaya Medical Centre
Wt	:	Body Weight
α	:	y-intercept
β	:	slope of line / regression coefficient
δ^2_E	:	Measurement Error
δ^2_S	:	Subject Variability
s^2_{YX}	:	Residual mean square
μ	:	Mean
δ	:	Variance
δ^2	:	standard deviation
^0C	:	degrees Celcius

List of Appendices

CHAPTER 1: INTRODUCTION

This thesis focuses on the application of statistical methods used to analyse continuous data in a method comparison study or a validation study in medicine. This chapter outlines the problem statement, the significance of this study and also its contribution to the medical field. It summarises the flow of this study and the structure of this thesis.

1.1 Study Background

In medicine, accurate measurement of clinical values is vital, either at the stage of health screening, diagnosing cases, or making prognosis. For example, accurate measurement of blood pressure, heart rate and oxygen level is crucial for monitoring patients under general anaesthetia in surgery. Inaccurate measurement of these variables will result in inappropriate management of the patient, thus putting the patient's life at risk.

Most of important variables measured in medicine are in numerical forms or continuous in nature, such as blood pressure, glucose level, oxygen level, weight, height, body temperature, creatinine level, albumin level, white cell count, platelet count, haemoglobin level, and many other clinical values. There are numerous instruments or machines that have been invented for the purpose of measuring various variables. Some measurements are obtained by using invasive techniques and expensive procedures. Consequently, new instruments and tests are constantly being developed and fashioned to provide complementary meaningful information to the search for information, with the aim of providing cheaper, non-invasive, more convenient and safe methods. Whether a test's outcome can provide trustworthy judgements or decisions depends particularly on the measurement quality of the test (Portney & Watkins, 2000).

When a new method of measurement or instrument is invented, the quality of the instrument has to be assessed. We want to know by how much the value of measurements obtained using new method differs from the old method, or from the gold standard. Information provided by any clinical instrument cannot be trusted and licitly used in any judgement and decision making process if the measurement quality has not been evaluated. This is where a method comparison study or a validation study comes into medicine.

Clinimetric properties indicating that the test is reliable and valid should be considered as fundamental for determining the measurement quality of any test (Feinstein, 1987). In general, clinimetric refers to the development of methodological and statistical methods applicable in clinical medicine in order to assign numbers or scores to observable clinical events (de Vet, Terwee, & Bouter, 2003a, 2003b).

1.1.1 Validity

An instrument is considered to be valid if it measures what it is intended to measure (de Vet et al., 2003b). The term "validity" actually has a wide range of classification and definition. In clinical research, current and accepted validity concepts include *criterion* validity, *construct* validity, and *content* validity, the first two being the most relevant for performance-based tests (Streiner & Norman, 2003).

Criterion validity is used to examine the extent to which a measurement instrument provides the same results as the gold standard (Streiner & Norman, 2003). This type of validity is the most powerful in terms of its usefulness, and is divided into two types: *concurrent* validity and *predictive* validity. Of these, *concurrent validity* is the most used method. This is when we are trying to compare a new measurement tool with the criterion measure, both of which are given at the same time (Haynes, Richard, & Kubany, 1995; Streiner & Norman, 2003). The new tool is usually simpler, cheaper

or less invasive compared to the standard or currently used tools. In contrast, in *predictive validity* the criterion will not be available until sometime in the future. When no gold standard is available, the common alternative is to use an accepted and well-grounded reference test to relate to the evaluated test (Baxter, 2005; Lambert, Gisel, & Wood-Dauphinee, 2002). Generally, this form of validity is used in developing instruments that allow us to get earlier answers, or to give earlier predictions than current instruments can provide (Streiner & Norman, 2003).

Construct validity refers to the degree to which a test measures a hypothetical, nonobservable construct, and this validity can be established by relating the test to outcomes of other instruments (Portney & Watkins, 2000; Streiner & Norman, 2003). It is used when we dealing with more abstract variables or factors that cannot be measured directly for example level of anxiety and pain (Streiner & Norman, 2003). We cannot see or directly measure anxiety, but we can observe other factors related to anxiety (according to theory) such as sweaty palm and tachycardia. The proposed underlying factors are referred to as *hypothetical construct* or simply known as *constructs* (Streiner & Norman, 2003). So, *construct validity* is the next best option in the absence of an acceptable gold standard. The measurement of instrument under study will be compared with other instruments that claim to measure the same construct (Streiner & Norman, 2003).

Content validity is a closely related concept, consisting of a judgement whether the instrument samples all the relevant or important content or domains (Innes & Straker, 1999; Streiner & Norman, 2003). Content validity can be claimed when a test logically and obviously measures what it purposes to measure (Haynes et al., 1995; Streiner & Norman, 2003). The relationship between the phenomenon being measured and the test score(s) is determined by a panel of experts or researchers (Haynes et al., 1995).

1.1.2 Reproducibility

Another approach in assessing the quality of measurement instrument is to assess the *reproducibility* of the instrument. This is when we are interested to know whether the new instrument is able to produce similar values as that predicted by the old or standard instrument. In the literature, terms reproducibility is often used interchangeably with the reliability, repeatability, consistency, agreement and stability (Innes & Straker, 1999). Recently, de Vet advocated that reproducibility is the proper term to use in clinical research, making the distinction between two aspects that are important for clinical interpretation: reliability and agreement (de Vet et al., 2003b; de Vet, Terwee, Knol, & Bouter, 2006).

1.1.2.1 Agreement

Agreement assesses how close the results of repeated measurements are to the "true value" or the criterion value (de Vet et al., 2006). So, agreement actually concerns accuracy or validity; more specifically, concurrent validity.

An instrument with good agreement will be able to produce accurate repeated measurements in the same person (de Vet et al., 2006). Thus, agreement parameters are important in instruments that are used for evaluative purposes. In evaluative measurement instruments, the variability between individuals in a population is not important, in comparison to the variability within an individual (de Vet et al., 2006). This is because, in some clinical settings, we want to detect differences or changes within the same individual, and not how much difference is the individual's value compared to another person's, or with the population. For example, in antenatal clinics we are interested in the weight gain of a mother throughout her pregnancy, and not how much her weight differs from the others'.

Agreement parameters estimate the *measurement error* in repeated measurements. When the measurement error is large, small changes cannot be distinguished from the measurement error (de Vet et al., 2006). The smaller the measurement error, the smaller the changes that can be detected beyond the measurement error, and the more appropriate the instrument is for evaluative purposes. Thus, for an instrument to be used to evaluate changes over time, such as changes in blood pressure after receiving antihypertensive therapy, it is important for us to ensure the agreement or the accuracy of the instrument.

1.1.2.2 Reliability

Reliability measures the extent to which test results can be replicated (de Vet et al., 2006). For example, if we measure body weight using a scale five times, ideally all five measurements should be the same. Reliability is concerned with precision. It also represents the extent to which individuals can be distinguished from each other, despite the variability of repeated measurements in one person or subject (i.e. measurement error) (de Vet et al., 2006).

In contrast with agreement, reliability measures the variability between people or subjects. This measurement tells us how well the measured value in one person can be distinguished from another (de Vet et al., 2006). Thus, reliability parameters are important when measurement instruments are used for discriminative purposes; for example, to decide whether a certain value is normal or abnormal, and when the measurement from the instrument is involved in important decisions, such as whether treatment is required or not.

In clinical practice, the cut-off for normal and abnormal values is usually well established by clinical guidelines, which are produced based on extensive reviews of available evidence. Reliable instruments should be able to provide values that will allow

doctors or clinicians to distinguish whether their patients are in the normal or abnormal group. For instance, if we take the blood pressure of one patient five times, all the values should be almost the same, and the values should give us an idea whether the patient's blood pressure is normal or not.

An acceptable range of reliability will vary depending on the circumstances (Streiner & Norman, 2003). For example, if repeated measurements of a weighing scale are found to vary around the "true" weight by 0.5kg, the reliability of this weighing scale would be acceptable if the measurements are only to be done on an adult population, but not reliable when used to weigh newborn babies in the hospital. This is because differences of 0.5kg in weight in an adult represents only a very small percentage of an adult body weight, and will not affect him or her clinically. In contrast, a difference of 0.5kg represents a large proportion of body weight for a newborn baby.

1.1.3 Agreement versus Reliability

To illustrate the concept of agreement and reliability in more simple language, imagine if we have three target boards (see Figure 1.1) that show the results of five repeated measurements of body weight of the same person, using three different scales (A, B and C). Figure 1.1(a) shows that after taking five measurements using scale A, the results of the measurements are scattered all over the target board. This suggest that the measurements are not near each other (poor reliability), and are not near their intended target or true value (poor agreement).

Figure 1.1(b) shows that all five measurements from scale B appear in more or less the same location on the target board, but not in the centre of the target board. This suggests that five different measurements were almost the same (good reliability), but they did not hit the intended target (poor agreement). Figure 1.1(c) shows that all five

measurements from scale C are close to each other (good reliability), and hit the centre of the target board (good agreement).

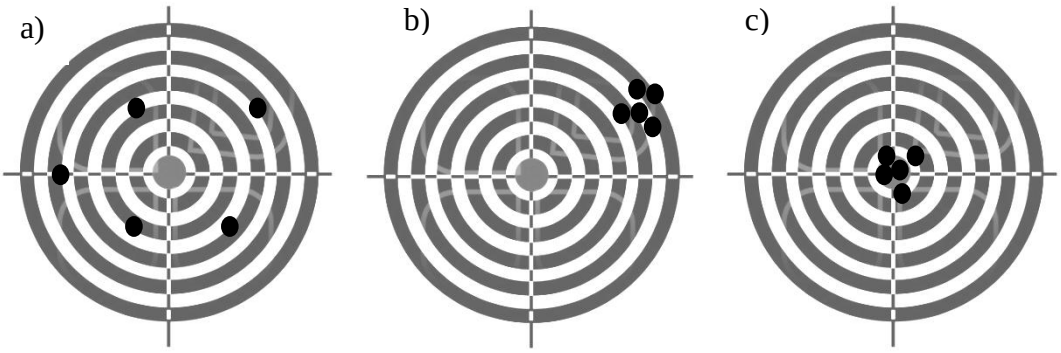


Figure 1.1: Results of measurements of body weight using three different scales A, B and C.

In most clinical situations, we use the same instrument to evaluate changes over time and also to differentiate values from the normal or abnormal cut-off point (which is usually derived from population-based studies). One of the examples of this situation is in the screening of hypertension cases, and the assessment of reduction of blood pressure post-treatment, in a clinic or health centre. Both blood pressure measurements are performed using the same blood pressure machine, or sphygmomanometer.

So, agreement and reliability parameters are equally important in determining the quality of instruments. In fact, it is difficult to be certain about the agreement of an instrument if the instrument is not reliable. Similarly, a precise instrument or instrument with good reliability will not necessarily measure the "true" value. Therefore, when comparing two instruments, or methods of measurement, we should consider assessing the *repeatability* of the instrument, which covers both agreement (accuracy) and reliability (precision).

1.2 Problem Statement

1.2.1 Inappropriate Application of Statistical Method

Thousands of validation studies have been conducted in the past. Various statistical tests have been used to test for agreement and reliability (Altman & Bland, 1983; Bahareh, Saeed, Ramin, & Bagher, 2008; Bruton, Conway, & Holgate, 2000). Some of the methods that were used were inappropriate. Correlation coefficient (r), coefficient of determination (r^2), regression coefficient, and means comparison have been shown to be inappropriate for the analysis in method comparison study. This has been discussed by Altman and Bland, since the 1980s (Altman & Bland, 1983), and also by Daly and Bourke (2000), and there is little argument about this in the literature (Daly & Bourke, 2000). Reasons for why those methods are inappropriate for the analysis in method comparison study will be discussed in detail in the next chapter (Chapter 2).

One example of the inappropriate application of statistical methods in method comparison study is in the study to explore the suitability of existing formulas to estimate the body surface area (BSA) of newborns (Ahn & Garruto, 2008). The authors compared different methods of estimation of body surface area in newborn, and used correlation coefficient to determine the agreement of those methods (Ahn & Garruto, 2008). In one of their results, the authors described that the method of estimating body surface area (BSA) using the BSA-Meban was most similar to the BSA-Mean, by having a mathematically perfect correlation with r = 1.00 ($p < 0.001$) (Ahn & Garruto, 2008). However, their conclusion was obviously inappropriate because the correlation coefficient only measures linear relationship, and does not suggest that the two methods give similar results.

Another example of the inappropriate application of the Pearson correlation coefficient was demonstrated in a recent study conducted in Greece (Milias, Antonopoulou, & Anthanasopoulos, 2008). The authors aimed to assess the validity of a

new motorised isometric dynamometer for measuring strength characteristics of elbow flexor muscles. They set the criteria of the Pearson correlation coefficient's (r) values > 0.97 to demonstrate that high agreement occurred between measures, and with r = 0.986, they concluded that the new dynamometer was accurate (Milias et al., 2008).

The use of inappropriate methods for the assessment of agreement and reliability will, undoubtedly, result in an inappropriate interpretation of the results and conclusions on the quality of an instrument. Consequently, this might result in the application of invalid equipment in medical practice, and will jeopardise the quality of care given to patients.

Altman and Bland proposed a method for agreement analysis in their original 1983 article (Altman & Bland, 1983). Later, they drew the attention of the medical professionals to this area in an article in *The Lancet* (Bland & Altman, 1986). Since then this article (Bland & Altman, 1986) has been cited in the literature more than 18,000 times (Bland & Altman, 2012). As a result of its high citation, the Bland-Altman method (Bland-Altman Plot and Limits of Agreement) is thought to be the most popular method in method comparison study. The popularity of the Bland-Altman method was thought, owing to its simplicity, practicality and ability to detect bias, when compared to other methods (Bahareh et al., 2008).

The issue of which method is the best is still debatable, and almost all methods have been criticised, especially for the agreement study. Even the Bland-Altman method has been criticised. Hopkins (2004) demonstrated that the Bland-Altman plot indicates, incorrectly, that there is systematic bias in the relationship between two measures (Hopkins, 2004). Details of this bias will be discussed in Chapter 2.

The question thus arises, how do researchers make their choices on which statistical methods to use? In fact, there is no clear guideline or recommendation for

researchers, especially for the clinician, on which is the best statistical method for analysis in method comparison study.

1.2.2 Application of Multiple Methods

The application of multiple or a combination of methods, particularly in the assessment of agreement, suggests that there is no consensus among researchers on which method is the best statistical method for measuring agreement. One example of the multiple application of method is in one study that testing the accuracy of peak flow meters (Nazir et al., 2005). In this study, the authors applied three statistical methods (Pearson's correlation coefficient, comparing mean (significant test), and the Bland Altman method) to assess for agreement of peak flow meters (Nazir et al., 2005).

A strong reason for using multiple methods in assessing agreement and reliability is that each statistical method has its strengths and weaknesses. The usage of multiple methods in method comparison studies has the advantage of compensating for the limitations of any one single method (Bruton et al., 2000; Luiz & Szklo, 2005).

1.2.3 Need for Guidelines or Recommendation

The question about the appropriateness of statistical analysis application in method comparison study suggests that there is a need for recommendation or guides for the medical professional on which statistical method is the best for measuring agreement and reliability.

The medical field is already full of complexity. Therefore a simple statistical method with less calculation and simple interpretation is preferable. A statistical method with simple calculation and interpretation will not only help to improve the understanding of the application of method, but will also reduce the errors that can be

made by the researcher. Nonetheless, the ability of the method in detecting bias is still the priority.

1.2.4 Need for Research

A review of various methods used in measuring agreement and reliability is required before any recommendation can be made. A comparison of the strengths and weaknesses of these methods can help to assess the best way of analysing data in a method comparison study. We need to see if there is any single method that is competent enough to detect agreement and reliability, or if the application of multiple methods is really necessary. If the application of multiple methods is needed, it is also important to identify the combination of which methods are the best for testing agreement and reliability.

1.3 Study Objectives

1.3.1 General Objective:

To compare different statistical methods of assessing agreement and reliability of medical instruments that measures the same continuous outcome.

1.3.2 Specific Objectives:

1. To perform a separate systematic review to identify the most commonly used statistical methods to assess agreement and reliability in medicine.
2. To compare most commonly used statistical methods in the analysis of agreement and reliability:
    - To determine which method is able to detect bias correctly
    - To determine how proportion and pattern of bias affect the prediction
    - To see the effect of sample size on the prediction of agreement/reliability
3. To make a recommendation on the most appropriate method to assess agreement and reliability

1.4 Significance of Study

1.4.1 Evidence-Based Medicine

The practice of Evidence-Based Medicine (EBM) has been promoted to ensure the best quality of care is given to the patient. One example is in the treatment of hypertension. According to the most recent National Institute of Clinical Excellence (NICE) *Clinical Guidelines on Hypertension* (NICE, 2011), antihypertensive drug treatment should be offered to people of any age with stage 2 hypertension. Stage 2 hypertension is defined as a patient with blood pressure of 160/100 mmHg or higher, and whose subsequent ambulatory blood pressure monitoring (ABPM), daytime average or home blood pressure monitoring (HBPM) average blood pressure, is 150/95 mmHg or higher (NICE, 2011).

The recommendation from the guidelines was derived from the views of experts, patients, carers and industry, and includes the best available evidence (from research) (NICE, 2011). Without doubt, researchers must have used some instrument to measure blood pressure in the process of producing evidence. However, which instrument was used in their studies: the automatic blood pressure machine or manual sphygmomanometer? Were these machines validated, and if the machines were validated, which statistical method was used? If the instruments used were not validated, or were validated using inappropriate statistical methods, we can actually question the quality of the evidence from such studies. A lack of precision and validity of an instrument in research may result in invalid evidence. The main goal of research, especially in epidemiological studies, is about applying the evidence to the population for practice. Appropriate statistical analysis is actually the “root” of Evidence-Based Medicine.

1.4.2 Patient Care

In clinical situations, the duty of a doctor is to provide the best care or treatment for their patients. Most of the time, doctors have to decide what is the best available option for their patients. In some cases, this may involve life and death decisions; for example, deciding to thrombolyse patient with myocardial infarction in an Accident and Emergency department. Doctors have to assess a patient thoroughly and, assisted by information from some medical equipments such as electrocardiogram (ECG) and blood pressure machines, before the decision to thrombolyse the patient can be made.

In 2009, a study to assess the accuracy and precision of five currently available blood glucose meters in South Africa was conducted (Essack et al., 2009). The study compared five glucometers that utilise different analytical techniques (reflectometry or amperometry), and all the glucometers were calibrated (Essack et al., 2009). The authors found that although all the devices showed satisfactory precision, there was substantial discordance when their results were compared to a laboratory reference (Essack et al., 2009). Only three out of the five glucometers fulfilled the criteria suggested by the International Standardisation Organisation. All meters demonstrated significant deviation from the American Diabetes Association guidelines, as more than 60% of the measurements exceeded the recommended percentage of deviation (Essack et al., 2009).

It is well-known that both type 1 and type 2 diabetes show a direct relationship between the degree of glucose control and the risk of systemic complications (M. Cohen, Boyle, Delaney, & Shaw, 2006). Many clinical organisations such as the American Diabetes Association promote the self-monitoring of blood glucose, because it allows diabetic patients to achieve and maintain specific glycaemic goals (M. Cohen et al., 2006). The variability observed with the accuracy of glucometers can impact

patient care in different settings, some of which include the diabetic patient on insulin in a home care or a clinical setting. Most of the time, glucose determinations and insulin adjustments are made according to glucometer readings. Inaccuracies can lead to misclassification of hypoglycaemic or hyperglycaemic episodes. It is, therefore, imperative that glucometer values are accurate and precise. Otherwise, a failure in this regard may lead to critical medical errors.

The variation amongst these glucometers found in the study (Essack et al., 2009) were probably a result of the improper evaluation of the glucometer in the validation study. This suggests that there is a necessity for proper evaluation, and it is important to be sure that appropriate statistical methods for the validation of the instrument has been used in any research or clinical situation.

If an instrument is not valid or reliable, it may lead to inappropriate conclusions. This will result in inaccuracy of prediction or diagnosis, and inappropriate management or treatment, which will definitely affect the quality of care given to a patient and, most importantly, inappropriate treatment might put the patient's life at risk. Poor quality of care will also jeopardise the doctor-patient relationship. Inaccurate measurements cannot be used as an excuse for making any mistake in the management of patients. Therefore it is vital to ensure the validity of an instrument, and appropriate statistical methods should be applied in a validation study. In other words, appropriate statistical methods should be used when testing agreement and reliability of an instrument.

1.5 Outline of Study

In general, this study focuses on two areas of analysis in method comparison study, which are agreement and reliability. To achieve the objectives of this study, this study was planned and conducted in five stages. The first stage is review of the literature, followed by a data collection and data analysis. Stage four is the synthesis of results.

The final stage is the discussion of findings, plus conclusions and recommendations. Stage 1 and stage 2 were conducted concurrently. Summary of the study outline is displayed in Figure 1.2.

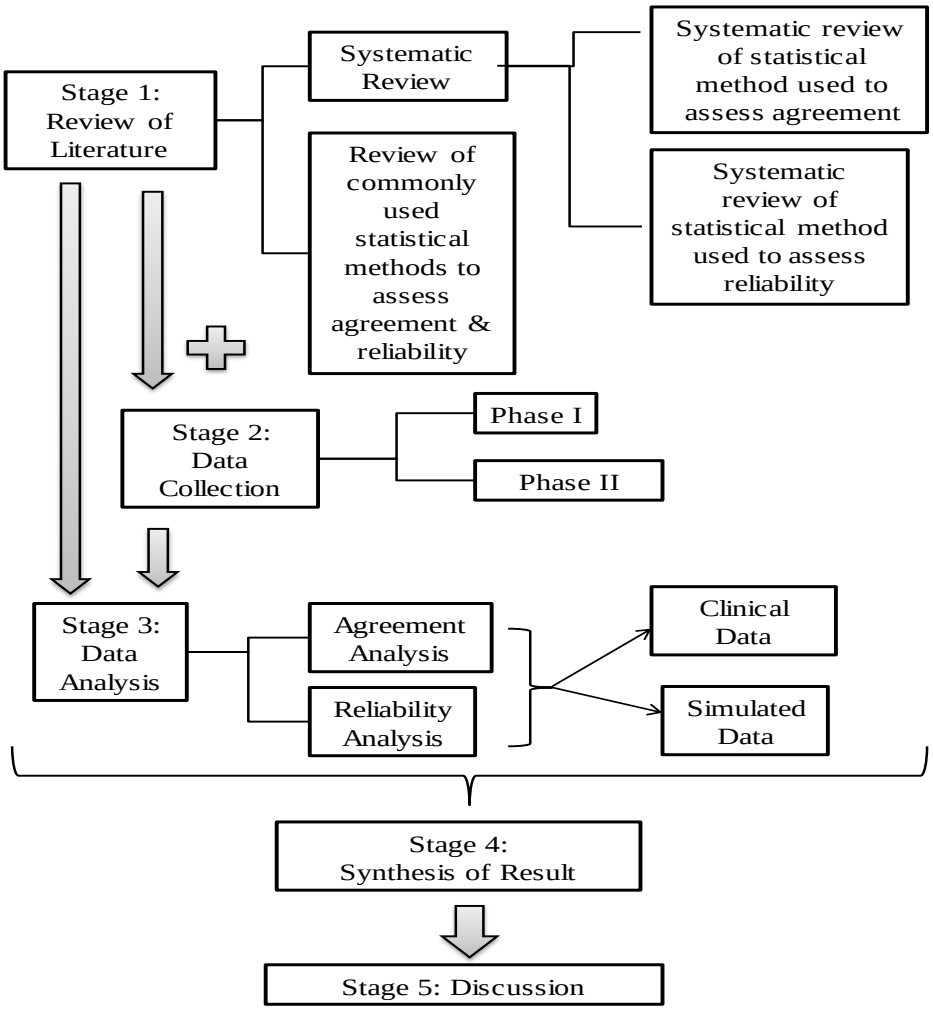


Figure 1.2: Outline of study

Stage 1: Data Collection

Data collection was performed in two phases due to limited resources and manpower. Phase I and phase II were conducted in different location and population. A total of eight variables were collected, to be used in the analysis of agreement and reliability.

Phase I was conducted between May and August 2009. Data were collected from a public university health screening program and a community health screening program run in a small village. Variables collected during this phase includes: blood glucose level; systolic blood pressure (SBP); diastolic blood pressure (DBP); and heart rate (HR).

Phase II was conducted between October 2009 and March 2010. Data were collected from a public university quit smoking clinic program and a community health screening program in a large shopping complex located in a city. Variables collected during this phase were: body weight; body temperature; peak expiratory flow rate (PEFR); and carbon monoxide level (CO). Details of the study population, data collection and data entry are described in Chapter 3. Results for the data collection are presented in Chapter 4.

Stage 2: Review of Literature

The literature review section is divided into two: the first part is a systematic review of the statistical methods used to assess agreement and reliability of medical instruments; the second part is the review of each statistical method found in the systematic reviews.

1.5.2.1 Systematic review

The purpose of the review is to identify the most commonly used method, to assess agreement and reliability of medical instruments measuring continuous variables. Two separate systematic reviews were performed.

The first review was conducted for the agreement study. The search and review of articles published between January 2007 and December 2009 was performed between March and April 2010, and analysis of the review was completed in June 2010. The search and review of articles for the second systematic review (i.e. the reliability study), was conducted between June and July 2010. Similar to the first systematic review, only articles published between January 2007 and December 2009 were included in this review. The analysis of the systematic review was completed in August 2010. A few issues related to the analysis in method comparison studies were found in these reviews. Findings from these reviews were presented at three international conferences, which resulted in two conference proceedings (Rafdzah Zaki, Bulgiba, Ismail, & Ismail, 2010; Rafdzah Zaki, Bulgiba, Nordin, Ismail, & Ismail, 2010), one full conference paper (Rafdzah Zaki, Nordin, Bulgiba, & Ismail, 2010), and one journal publication (R. Zaki, Bulgiba, Ismail, & Ismail, 2012). Details of both systematic reviews are discussed in Chapter 2.

1.5.2.2 Review of each statistical method

The second part of the literature review involves the review of the most commonly used statistical methods to assess agreement and reliability. This is a thorough review of the literature on the theoretical concept of each statistical method, and the suitability of each method for the analysis of agreement or reliability. This review is presented in Chapter 2.

Stage 3: Data Analysis

Data analysis was divided into two study areas: the analysis of agreement; and the analysis of reliability. Each substudy involved the analysis of original clinical data and simulated data. The analysis for the agreement study was performed between January and May 2011. The analysis for the reliability study was performed between June and September 2011. Results of the analyses are presented in Chapter 4.

Learning about and familiarity with all the software used in this project also formed part of the preparation for the analysis. The statistical software packages used in this study were: SPSS 17.0, GraphPad Prism 5.02, Matlab 7.8 and MedCalc 12.1.3. The greatest challenge was learning the Matlab software.

Stage 4: Synthesis of Results

Stage four of this study is the synthesis of the results; the evaluation of the results before a detailed discussion of the findings in this study. There were few unexpected findings in the analysis, which resulted in extended analysis in the analysis of agreement and reliability. Extended analysis for both agreement and reliability study was performed between December 2011 and March 2012, and the results are presented in Chapter 4. A few papers for publication were prepared during this stage.

Stage 5: Discussion

Finally, after taking into account all the information from the literature review and findings from the analysis, a discussion on the topic of research and recommendations were made. The discussions of findings for both agreement and reliability studies are presented in Chapter 5. The conclusion and recommendations are presented in Chapter 6.

1.6 Contribution of Study

This study has contributed to three important areas in the medical field. The first contribution is to the systematic review of statistical methods used to assess agreement and reliability of medical instruments measuring continuous variables. Both reviews are the first reviews ever of these topics. From the reviews, some issues related to method comparison studies were discussed, including the application of inappropriate statistical methods, and the importance of education in method comparison studies among medical professional. The second contribution is the comparison of different statistical methods from an extensive analysis of real clinical and simulated data. Finally, the most important contribution of this study is the recommendations on a few issues related to method comparison study. Apart from these three main contributions to the medical field, this study has also made a social contribution to the local community in the form of free health screening and consultation.

1.6.1 Systematic Review of Statistical Methods Used

One of the important contributions of this study is the systematic review of the statistical methods used to assess agreement and reliability. Two separate systematic reviews were performed as part of this study. Each review identifies the most common statistical methods used to assess agreement and reliability, in recent validation studies conducted in medicine. Both reviews are the first systematic reviews of the topics. This provides evidence on the most popular statistical methods used in the analysis of validation study in medicine, which reflects the current knowledge of statistical methods among medical researchers.

Both reviews also showed that there were inappropriate applications of statistical methods to assess agreement and reliability in recent studies. It is important for a clinician or medical researcher to be aware of this issue because it would be dangerous

if a misleading conclusion from inappropriate statistical analysis led to the application of inaccurate instruments in clinical practice. This also suggests that method in validation study is an important area that should be explored by medical professionals, and should not be neglected in medical education. The issue of inappropriate analysis in method comparison study should also be highlighted so that the same mistakes are not repeated by future researchers.

1.6.2 Comparison of Statistical Methods

This study provides extensive analysis of the most commonly used statistical methods in the analysis of agreement and reliability, using both clinical datasets and simulated datasets. The results of the extensive analysis allow a comparison of the strengths and weaknesses of each method. This study also presents the explanation of the theoretical concepts behind the most commonly used statistical methods in a method comparison study, in plain and "non-technical" language for the benefit of the medical professional. In addition to comparing the most commonly used statistical methods in method comparison study, this study also compares the potential of the comparing slopes and y-intercepts and the proposed method of agreement model in the analysis of agreement.

1.6.3 Recommendation related to Method Comparison Studies

The final contribution of this study is recommendations on a few issues related to the method comparison study, including how to conduct a method comparison study, and the importance of educating medical researchers and clinicians on method comparison study. The most important outcome from this study is recommendations on the most appropriate way to analyse data in agreement and reliability studies. Hope these recommendations able to solve the problem of inappropriate statistical methods in the

analysis of agreement and reliability. These recommendations can be found in Chapter 6. The findings and recommendations made in this study will not only help medical professionals in conducting method comparison studies, but will also help them in appraising other people's studies (on deciding the validity and precision of certain medical instruments).

1.6.4 Social Contribution

The health screenings sessions conducted by the researcher both in the university and community setting during data collection has contributed to the general health and well-being of the local community. Free health screening and consultation during the sessions would definitely benefit all participants. The free health screening was conducted by the researcher who is a qualified medical doctor. The opportunity created by the researcher was really appreciated by all the participants, especially the villagers.

Most of the members in the village do not attend any regular health screening, and only seek medical help when they are really unwell. There were a few health problems detected among the participants during the screening session, including high blood pressure and a high blood sugar level. The opportunity was also used by the villagers to have a general medical consultation with the researcher.

1.7 Thesis Structure

This thesis is divided into six chapters. Chapter 1 is the introductory chapter, summarises the work contributed by the author, and outlines the general layout of this thesis. Apart from the aim and objectives of this study, this chapter also presents an introduction to the validation study in medicine, the concept of agreement and reliability, and the importance of this study area in medicine.

Chapter 2 presents two systematic reviews that were conducted in this project. First, is a systematic review of methods to assess agreement in medicine, and second, is a systematic review of statistical methods used to assess reliability in medicine. Commonly used statistical methods to assess agreement and reliability are presented, and each of these methods is discussed in terms of its method of calculation, assumptions and interpretation of the results. A proposed simple method of assessing agreement is also described in this chapter.

Chapter 3 is the methodology of this study. Starting with ethical clearance and funding application, this chapter describes the major works involve in producing this project, including details of work done in data collection, data management and data processing for use in the research project. Problems encountered in the data collection, data management and processing are discussed, and solutions to these problems are offered.

Chapter 4 presents the results of the data analysis. This chapter compares the main statistical methods used to assess agreement and reliability. Analysis of agreement and reliability are discussed separately under different subtopics. The performance of each statistical method is compared, according to the effect of sample size, proportion of bias in the dataset, consistency of error in the data, and the range of the dataset.

Chapter 5 is a discussion of the comparison of all the statistical methods used to assess agreement and reliability. The advantages and disadvantages of each statistical method used to assess agreement and reliability are also discussed in this chapter.

Finally, Chapter 6 concludes chapter and rounds up the thesis. Recommendations based on the work done in this thesis are presented. Furthermore, general aspects and recommendations for future research are also proposed.

1.8 Summary of Chapter 1

This chapter contains the introduction to this study. This chapter described the background to this study, and introduced the topic of validation study in medicine, including the "agreement" and "reliability" parameters. This chapter also discussed problems in validation study and its implication for medical practices. Besides highlighting the importance of appropriate statistical analysis in medical research, this chapter also presented the importance of this area of study in medicine, and discussed the topic of agreement and reliability, as well as the objectives of the study. The outline of the research approach and the structure of the thesis were explained, and contributions of this project to the medical field are detailed. There are three main contributions of this thesis. The first is the systematic reviews of statistical methods used to assess agreement and reliability in medicine, which highlight the inappropriate application of statistical methods in measuring agreement and reliability. The second contribution is a comparison of the most commonly used statistical methods to assess agreement and reliability, based on extensive analysis of real clinical data and simulated data. Finally, the third contains recommendations on a few issues in method comparison study, especially recommendations on what is the best way to assess agreement and reliability, which should be able to guide the medical researchers and medical practitioners when conducting research, appraising other people's studies, and also in their daily clinical practice.

CHAPTER 2: REVIEW OF LITERATURE

2.1 Introduction

This chapter presents reviews and appraises the most recent, related literature to the area of research. Two separate systematic literature searches for method comparison studies are presented in this chapter: one for agreement (Section 2.2) and another for reliability studies (Section 2.3). These reviews look at previous validation studies in medicine. The main objective is to identify the most commonly used methods to assess agreement and reliability of medical instruments that measure continuous variables in recent studies. Details of the most commonly used statistical methods for agreement and reliability are discussed in each section, and covers the methods of calculation, assumption, interpretation of the results and suitability to assess agreement or reliability, where applicable.

Issues regarding the methods in the analysis of method comparison study are discussed in Section 2.4. Later in this chapter (Section 2.5), the theoretical concept of proposed statistical method for assessing agreement, based on analysis of linear regression lines, is briefly discussed. However, the suitability of this method will be tested later in data analysis (Chapter 4). Section 2.6 summarises recent evidence of the highlighted problems discussed in Chapter 1, and re-emphasises the importance of this research in medicine.

2.2 Methods of Measuring Agreement

2.2.1 Systematic Review of Methods Used to Assess Agreement

The purpose of this section is to review the statistical methods used to measure agreement of equipment measuring continuous variables in the medical literature. So

far, no other study has been designed specifically to look at this issue. This review aims to identify statistical methods used to assess the agreement of equipment measuring continuous variables in recent studies (in medicine). This will reflect the statistical knowledge of method comparison studies among medical researchers. Moreover, the proportion of various statistical methods found in this review will reflect the proportion of medical instruments that have been validated, using those particular statistical methods in current clinical practice. Therefore, this review only includes the most recent articles which were published from 2007 to 2009. An Internet-based search was used, and only full text articles were included in this review. Unpublished articles were not considered. This review follows the standards as suggested in the PRISMA statement (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009). The PRISMA checklist for this systematic review is attached as Appendix I.

2.2.1.1. Literature search and study selection

The search for articles was performed in January 2010, from five main medical databases (Medline [EBSCOhost], Ovid, PubMed, Scopus and Science Direct) for studies investigating the agreement of instruments or equipment in medicine, and published in journals between January 2007 and December 2009. Since the focus of this research is on the method of validation study for medical instrument with continuous variables, only studies that investigated the agreement of equipment measuring continuous variables were included in this review. A Boolean search was performed on each database using the search term: Agreement AND (validation OR "comparison study"). The search was limited to the medical field (including dentistry), studies involving human subjects, and articles written in English.

Table 2.1 presents the summary of the literature search. Initial search limitations were based on the search system of each database. All citations and abstracts were exported to the Endnote software, and then a search for duplicates was performed. Any studies with qualitative or categorical data, studies with different units of outcomes, and association studies were excluded. The study selection process is summarised in Figure 2.1.

Table 2.1: Systematic search of article

Database	Search limitation according to database	Search Term/strategy	Total hits	Total
Ovid	January 2007– December 2009 English Human Full text	#1 Agreement #2 Validation Study #3 Comparison Study **#1 AND (#2 OR #3)**	1,766 51 146	**1**
Scopus	Year 2007–2009 English Medicine (subject area) Article (document type)	#1 Agreement #2 Validation study #3 Comparison study **#1 AND (#2 OR #3)**	11,184 5,501 1,394	**558**
Medline (EBSCO host)	January 2007– December 2009 Full text	#1 Agreement #2 Validation study #3 Comparison study **#1 AND (#2 OR #3)**	128,133 4,257 1,968	**941**
Science Direct	Year 2007–2009 Medical & Dentistry (subject area) Journal article (document type)	Search term: Agreement AND (“validation study” OR “comparison study”)		**1,654**
PubMed	Year 2007–2009 English Human Full text	#1 Agreement #2 Validation study #3 Comparison study **#1 AND (#2 OR #3)**	11,008 622 299	**106**
		TOTAL		**3,260**

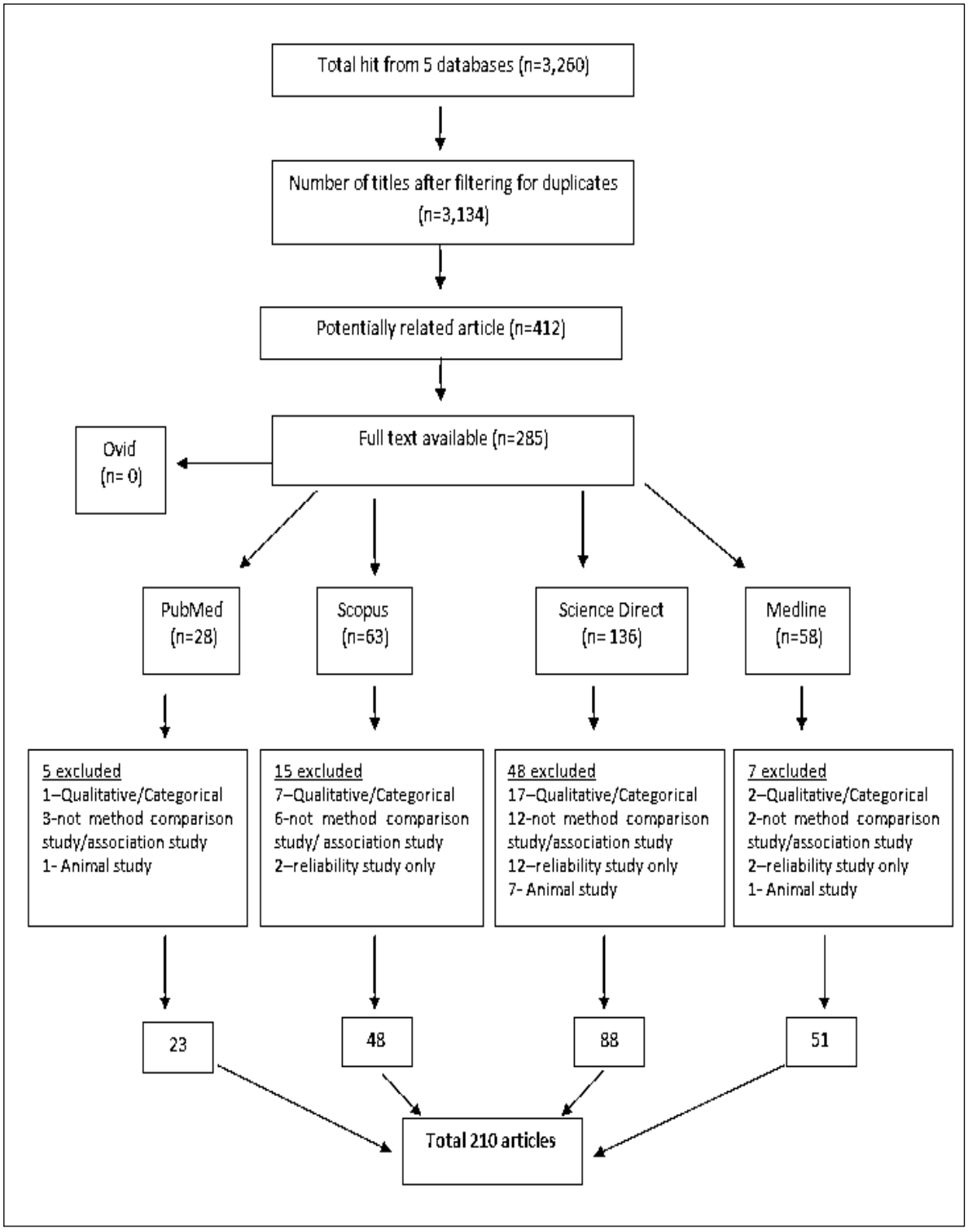


Figure 2.1: Flow chart of the final study selection

2.2.1.2 Data extraction and analysis

Information on the year of publication and journal types was extracted from each article. The journal types were divided into five areas: medicine (including emergency and critical care medicine); surgery; radiology; nutrition; and others. Information on the statistical methods used to assess agreement were determined according to the stated statistical analysis under the method section, and also by identifying which statistical method(s) influenced the author's conclusion on the agreement.

Data were analysed using SPSS 17.0 software. Descriptive analysis of the characteristic of studies and statistical methods used were performed. Univariate analysis was performed between statistical methods used and covariates (journal type, year of publication, and online databases), using Chi-square test and Fisher's exact test (where appropriate). Probability value (*p*-value) of <0.05 was considered to be significant.

2.2.1.3 Findings from the systematic review of agreement studies

a. Characteristics

Of the 210 articles reviewed, 70 were published in 2007, 70 in 2008, and 70 in 2009. Of these, 88 (42%) of the articles were obtained from the Science Direct database, 51 (24%) from the Medline database, 48 (23%) from the Scopus database, and 23 (11%) from the PubMed database. Most of the studies (72, or 34%) were published in medical journals, 30 (14%) in nutritional related journals, 29 (14%) in radiological journals, 28 (13%) in surgical journals, and the rest from other areas (such as public health, ophthalmology, biomedical, psychology, and dentistry).

b. Statistical Method used

Overall, 117 articles (56%) used a single method to assess agreement, while 93 articles (44%) used multiple (2 or more) methods. The five most popular statistical methods used to assess agreement in the 210 reviewed articles are summarised in Table 2.2. Most of the articles (178 articles or 85%) have used the Bland-Altman Limits of Agreement method to measure the agreement of equipments. Out of the 178 articles, 99 studies (56%) used the Bland-Altman method alone to assess agreement and the remainder (79, or 44%) used a combination of Bland-Altman method and another method. Only 62 or 30% of agreement studies also assessed reliability. The Bland-Altman method is popular for most specialties in medicine especially in radiology (see Table 2.3).

Twenty articles (10%) used clearly inappropriate methods in their study. These articles (Ahn & Garruto, 2008; Ahn et al., 2007; Allen, Wallace, Larson, Sheppard, & Liu, 2007; Anderst, Zauel, Bishop, Demps, & Tashman, 2009; Andrieux, Kilinc, Perrin, & Campos-Gimenez, 2008; Barthelemy, Gregor, Krejci, Wataha, & Bouillaguet, 2009; Camara et al., 2008; Chovel Cuervo, Sterling, Abreu Nicot, García Rodríguez, & Rodríguez García, 2008; Cuker et al., 2009; Di Noia & Contento, 2009; Hacihaliloglu, Abugharbieh, Hodgson, & Rohling, 2009; Hof et al., 2008; Jaffrin & Morel, 2008; Mündermann, Dyrby, & Andriacchi, 2008; Naidu, Panchik, & Chinchilli, 2009; Reis, Aniceto, Aguiar, Simao, & Segurado, 2007; Satia & Galanko, 2007; Satia, Watters, & Galanko, 2009; Shuaibi, Sevenhuysen, & House, 2008; Ten Boekel et al., 2007) used either the correlation coefficient, coefficient of determination, comparison of means, or a combination of these methods in the analysis of agreement. No significant association was found between the statistical methods used and the year of publication ($p = 0.62$), electronic database ($p = 0.06$), and type of journal ($p = 0.42$).

Table 2.2: Most popular statistical methods used to assess agreement in medicine

Statistical Method Used		Number of method used in all the 210 articles n (%)	Number of method used according to year of publication		
			2007 (n=70)	2008 (n=70)	2009 (n=70)
1. Bland-Altman method (limits of agreement)	Yes	178 (85%)	62 (89%)	59 (84%)	57 (81%)
	No	32 (15%)	8 (11%)	11 (16%)	13 (19%)
2. Correlation coefficient (*r*)	Yes	58 (28%)	21 (30%)	24 (34%)	13 (19%)
	No	152 (72%)	49 (70%)	46 (66%)	57 (81%)
3. Compare mean/ Significant test	Yes	38 (18%)	9 (13%)	15 (21%)	14 (20%)
	No	172 (82%)	61 (87%)	55 (79%)	56 (80%)
4. Compare slope and intercept	Yes	13 (6%)	7 (10%)	4 (6%)	2 (3%)
	No	197 (94%)	63 (90%)	66 (94%)	68 (97%)
5. Intra-class Correlation Coefficient	Yes	14 (7%)	1 (1%)	6 (9%)	7 (10%)
	No	196 (93%)	69 (99%)	64 (91%)	63 (90%)

Table 2.3: Top five statistical methods used to assess agreement according to area of specialty in medicine

Statistical Method Used	Number of article using the method, n (%)
Medicine (N = 29)	
1. Bland-Altman Limits of Agreement	24 (83%)
2. Correlation coefficient (r)	6 (21%)
3. Compare slope or/and intercept	4 (14%)
4. Intra-class Correlation Coefficient	3 (10%)
5. Compare mean/ Significant test	2 (7%)
Surgery (N = 25)	
1. Bland-Altman Limits of Agreement	21 (84%)
2. Correlation coefficient (r)	8 (32%)
3. Compare mean/ Significant test	5 (20%)
4. Intra-class Correlation Coefficient	4 (16%)
5. Percentage of error	1 (4%)
Radiology (N = 29)	
1. Bland-Altman Limits of Agreement	26 (90%)
2. Correlation coefficient (r)	6 (21%)
3. Compare mean/ Significant test	6 (21%)
4. Intra-class Correlation Coefficient	3 (10%)
5. Compare slope / intercept	2 (7%)
Nutrition (N = 30)	
1. Bland-Altman Limits of Agreement	25 (83%)
2. Correlation coefficient (*r*)	13 (43%)
3. Coefficient of determination (r^2)	4 (13%)
4. Compare mean/Significant test	4 (13%)
5. Compare slope or/and intercept	4 (13%)

N = Total number of studies retrieved for each specialty, n = number of studies, % = percentage

2.2.2 Review of Most Commonly Used Methods to Assess Agreement

2.2.2.1 Bland-Altman Limits of Agreement

In 1983, Bland and Altman introduced Limits of Agreement (LoA) to quantify agreement (Altman & Bland, 1983). Bland and Altman (Bland & Altman, 1987) stated that it is very unlikely for two different methods or instruments to be exactly in agreement, or give identical results for all individuals. However, what is important is how close the values obtained by the new method (predicted values) are to the gold standard method (actual values). This is because a very small difference in the predicted and the actual value will not have an effect on decisions of patient management (Bland & Altman, 1987). So they started with an estimation of the difference between measurements by two methods or instruments (Bland & Altman, 1987).

To construct Limits of Agreement, first we need to calculate the mean and standard deviation of these differences. The formula for Limits of Agreement (LoA) is given as (Bland & Altman, 1987):

$$\boldsymbol{LoA = mean\ difference \pm 1.96 \times (standard\ deviation\ of\ differences)}$$

So, 95% of differences should lie within these limits. To illustrate this, we can use the data from Table 2.4 (adapted from Table 12.5, *Interpretation and Uses of Medical Statistics*) (Daly & Bourke, 2000), which compared the values from the glucometer and laboratory. If we apply the data from Table 2.4, the first step of the analysis is to calculate the difference and mean. The mean difference for the data is -0.28mmol/l, and the standard deviation of difference is 0.27mmol/l. This makes the LoA = -0.81mmol/l to 0.26mmol/l.

Table 2.4: Hypothetical data of blood glucose level from a glucometer and laboratory.

Patient	Lab Value (L)	Glucometer (G)	G–L	Mean
1	10.20	10.20	0.00	10.20
2	8.20	8.00	-0.20	8.10
3	8.70	8.05	-0.65	8.38
4	9.60	9.70	0.10	9.65
5	9.60	9.05	-0.55	9.33
6	8.20	8.15	-0.05	8.18
7	9.40	8.80	-0.60	9.10
8	7.00	6.55	-0.45	6.78
9	6.60	6.55	-0.05	6.58
10	10.80	10.50	-0.30	10.65

a. Interpretation of Limits of Agreement

Limits of Agreement give us the range of how much one method is likely to differ from another. So it is all about the differences. If we are testing a new method B against the old method A, and the difference is calculated from A–B, then a positive value of limits of agreement means A>B, or new method B underestimates the new method A. If a negative value of limits of agreement means A<B, or the new method B overestimates the old method A. So, the result of Bland-Altman analysis between glucometer and laboratory values (Table 2.4) can be shown as follows:

Differences = Glucometer – Laboratory

Mean difference = -0.28 mmol/l

Limits of Agreement = -0.81mmol/l to 0.26mmol/l

This means that, on average, the glucometer measures 0.28 mmol/l less than the laboratory. Also, 95% of the time the glucometer reading will be somewhere between 0.81mmol/L below and 0.26mmol/l above the laboratory values.

b. Assumptions

The 95% Limits of Agreement is dependent on the assumptions that the mean and standard deviation of the differences are constant throughout the range of measurement, and the distribution of these differences follow approximately a normal distribution (Altman & Bland, 1983). It is important to check for these assumptions (Altman & Bland, 1983). Altman and Bland (1983) proposed a scatter plot of the differences of two measurements against the average of the two measurements, and a histogram of the differences, to check for these assumptions (Altman & Bland, 1983). Initially, the scatter plot is only to check the assumption and not the analysis of agreement, but then it becomes a graphical presentation of agreement (see Figure 2.2).

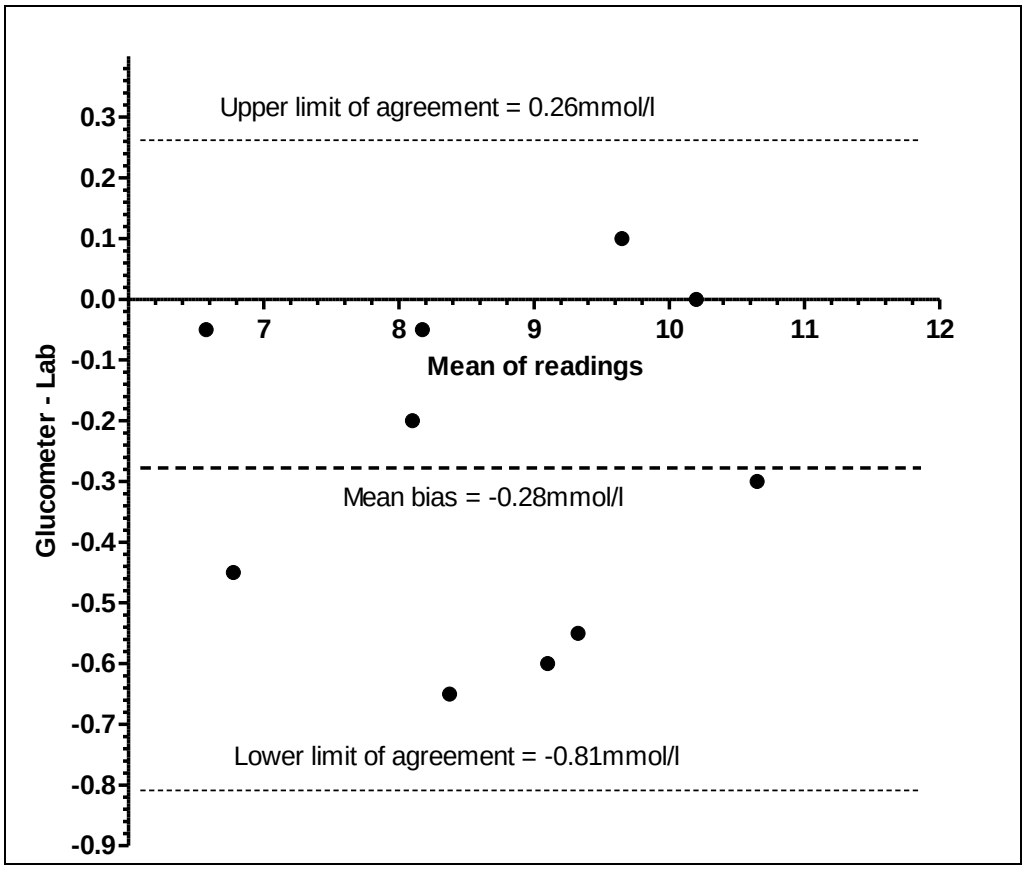


Figure 2.2: The Bland-Altman Plot

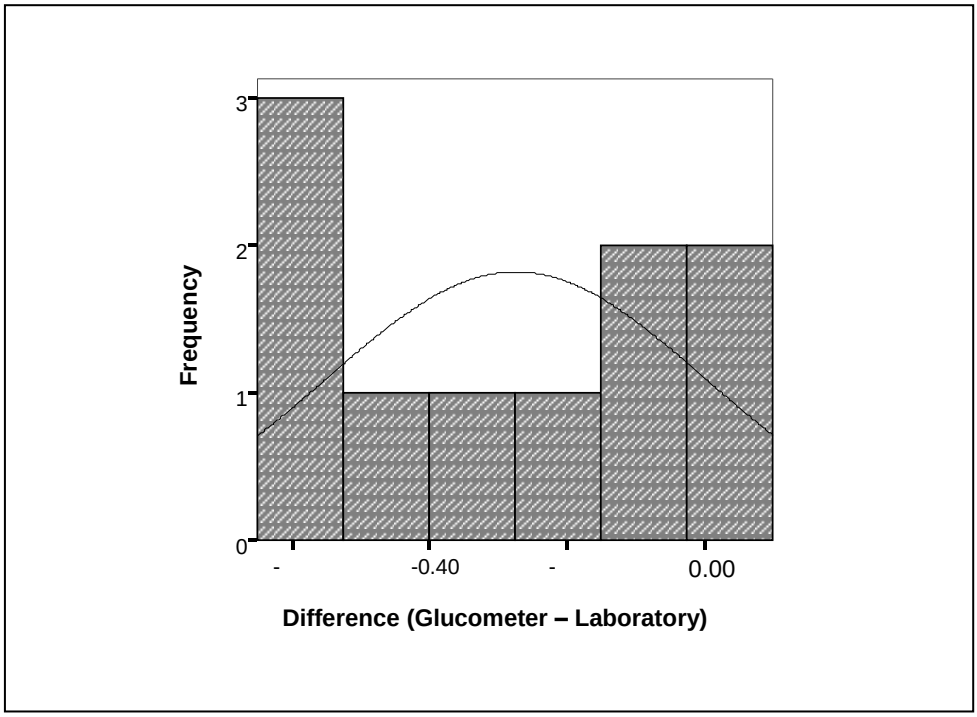


Figure 2.3: Distribution of Differences

From the histogram above (Figure 2.3), the distribution of differences is not normal. The assumption of the normality of differences for the data from Table 2.4 is, therefore not met. So the Limits of Agreement calculated previously is questionable. In the situation where normality is violated, Bland and Altman (Bland & Altman, 1986) proposed a logarithmic transformation of both measurements before analysis. The limits of agreement can be back-transformed to give limits for the ratio observations that lie above the line of equality (Bland & Altman, 1986). Bland and Altman do not recommend any other methods of transformation because the log transformation is the only transformation giving back-transformed differences, which are easy to interpret (Bland & Altman, 1986).

Table 2.5: Log transformation data

Log of Lab value (Log L)	Log of Glucometer (Log G)	Log G−Log L	Mean
1.01	1.01	0.00	1.01
0.91	0.90	0.01	0.91
0.94	0.91	0.03	0.92
0.98	0.99	-0.01	0.98
0.98	0.96	0.02	0.97
0.91	0.91	0.00	0.91
0.97	0.94	0.03	0.96
0.85	0.82	0.03	0.83
0.82	0.82	0.00	.82
1.03	1.02	0.01	1.03

Table 2.5 shows the transformed value of data from Table 2.4. The Bland-Altman Method analysis of the log transformed data is:

Bias (Limits of Agreement) = -0.01(- 0.04 to 0.02)

while the back-transformed of these values is:

Bias (Limits of Agreement) = 0.98 (0.91 to 1.05)

The antilog of the difference between two values on a log scale is a dimensionless ratio (Bland & Altman, 1986). The limits tell us that for about 95% of cases the measurement of glucometer will be between 0.91 and 1.05 times the laboratory value. Thus the glucometer measurement may differ from the laboratory measurement by 9% below to 5% above.

2.2.2.2 Correlation Coefficient

One of the favourite approaches in measuring agreement is to calculate the correlation coefficient (*r*) (Altman & Bland, 1983; Fay, 2005; Lee, Koh, & Ong, 1989). As found in the review earlier in this chapter, this method is the next most popular method after the Bland and Altman method, used to assess agreement. The first approach in this analysis is to make a scatter diagram, and then to calculate product-moment correlation coefficient (Fay, 2005). To calculate the product moment correlation coefficient (*r*), variables for each pair of measurements are labelled as X and Y. The formula for the correlation coefficient *r* is given as:

$$r = \frac{\sum \mathrm{XY} - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{N}\right)}}$$

If we use an example from data presented in Table 2.4 (to compare blood sugar levels from glucometer and laboratory values), the Pearson correlation coefficient (*r*) is 0.9798 with a 95% confidence interval of 0.9139 to 0.9954, and *p*-value <0.0001 (analysis using SPSS 17.0 software). The null hypothesis here is that the measurements of blood glucose level by the two methods (glucometer and laboratory) are not related linearly. With a very small *p*-value, we can reject this null hypothesis and propose an alternative hypothesis: there is a linear relationship between the measurements of glucose level by the two methods (glucometer and laboratory). Some people will interpret this as being that there is an agreement between the two instruments. This is another mistake conducted by many researchers (Altman & Bland, 1983).

Correlation is a measure of association, and only measures the strength of linear relationship (Fay, 2005). Strong correlation does not mean strong agreement. To demonstrate the inappropriate use of correlation, let's double the value of glucometer

from Table 2.4 so that it is obvious that there is no agreement between the glucometer and the laboratory value (see Table 2.6). Despite this, the correlation analysis of data from Table 2.6 will give exactly the same Pearson correlation coefficient (*r*) of 0.9798, with a similar 95% confidence interval (CI) of 0.9139 to 0.9954. Of course the two instruments (glucomoter and laboratory measurement) do not agree, but the correlation coefficient value is still very high, suggesting a strong correlation or association.

Table 2.6: Hypothetical data of blood glucose value

Lab Value	**Glucometer**	**Glucometer x2**
10.20	10.20	20.40
8.20	8.00	16.00
8.70	8.05	16.10
9.60	9.70	19.40
9.60	9.05	18.10
8.20	8.15	16.30
9.40	8.80	17.60
7.00	6.55	13.10
6.60	6.55	13.10
10.80	10.50	21.00

Agreement is assessing a different aspect of relationship between two measurements as compared to the correlation coefficient. The correlation coefficient reflects the noises and direction of a linear relationship (Bland, 1995; Lin, 2000). Perfect correlation occurs if all the points lie along any straight line (see Figure 2.4), and so data with poor agreement can produce a high or strong association (Bland & Altman, 1987). Furthermore, data covering an extensive (wide) range of values will appear to be more highly correlated than if it covers a narrow range (Bland & Altman, 1987). Therefore, it is clear that correlation is not an appropriate method for testing agreement.

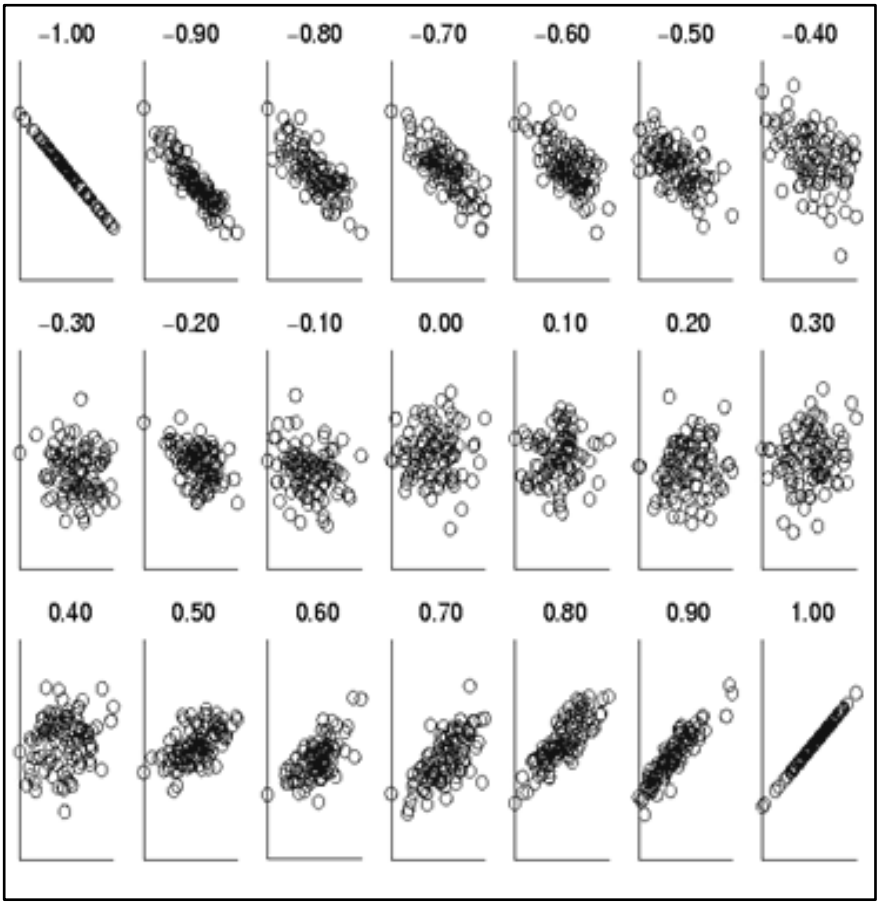


Figure 2.4: Correlation coefficient values, and the noises and direction of a linear relationship (figure adapted from Pennsylvania State University online course, 2013).

Some people use the *coefficient of determination* (r^2) parameter as a measure of agreement. One example of the application of this method is in a recent study (Hanbazaza & Mansoor, 2012) on the accuracy evaluation of point-of-care glucose analysers in the Saudi Arabian market. The authors compare the blood glucose readings from five different types of glucose analysers with the results from laboratory analysis. Their aim was to test the accuracy of the devices. In one of their results, the authors described that the Nova StatStrip device showed an excellent performance that almost agreed and correlated perfectly with the lab results, because the r^2=0.99 (Hanbazaza & Mansoor, 2012).

However, the use of the *coefficient of determination* (r^2) is inappropriate because r^2 is obtained from correlation coefficient *r*, which is a wrong method to measure agreement. *Coefficient of determination* (r^2) is used to state the proportion of variance in the dependent variables that is explained by the regression equation or model (Daly & Bourke, 2000). The more closely the points are dispersed around the regression line in the scatter diagram, the higher the proportion of variation explained by the regression line, thus the greater the value of r^2 (Fay, 2005). So it applies a similar concept to the correlation coefficient.

2.2.2.3 Comparing means

The third most popular method found in the systematic review (Section 2.2.1) is comparing means of readings from two instruments. In this method, the means of readings from two instruments are compared. The test of significance is then carried out to test the null hypothesis that there is no difference between the means of readings from the two instruments.

In assessing agreement, the same measurement of similar subjects will be taken using different instruments. Therefore a paired t-test is usually used to test the

hypothesis. Here, we want to know whether the difference observed is the true difference or has only occurred by chance when there was really no difference in the population. If the difference is truly occurring, and the null hypothesis is not true, then the alternative hypothesis must be true. So, in this case, the alternative hypothesis is that there is a significant difference between the mean of reading from the two instruments.

People have interpreted non-significance results to mean that there is not enough evidence to show that the two means differ (i.e. no differences), thus there is an agreement between the two groups, and vice versa. An example of this inappropriate approach is in a study conducted in Sweden on the assessment of left ventricular volumes, using simplified 3-D echocardiography and computed tomography (Mårtensson et al., 2008). However, the paired t-test with non-significant results does not indicate agreement. The reason for this is that the value of mean is affected by the value of each individual data, especially when there is an outlier. Distribution of differences between the instruments can lead to a difference in means being non-significant. It is possible that poor agreement between the two instruments can be hidden in the distribution of differences, and thus the two methods can appear to agree (Daly & Bourke, 2000). To illustrate this example, we have a hypothetical dataset comparing the measurements from standard instrument A, with the new instruments B and C (Table 2.7).

Table 2.7: Hypothetical dataset for instruments A, B and C

Patient	A	B	C
1	1	1	5
2	2	3	4
3	3	2	3
4	4	5	2
5	5	4	1

From the dataset (Table 2.7), it is obvious that the two new instruments (B and C) do not agree with the standard instrument A. The mean and standard deviation for the three groups are all the same: the mean is equal to 3.0 and standard deviation is equal to 1.58. If we compare the readings from instruments A and B, using a paired t-test, the results will be:

Mean Difference (Confidence Interval) = 0 (-1.24 to 1.24)

Standard Deviation of Differences = 1.0

p-value = 1.0

So, from this analysis we can conclude that there is no difference between the mean reading of instruments A and B. If we are saying that non-significant results indicate an agreement, this suggests that there is an agreement between instruments A and B. However, we know that this is not true. Similarly, the result will be not significant when we compare the mean reading of instruments A and C, where the results will be:

Mean Difference (Confidence Interval) = 0 (-3.93 to 3.93)

Standard Deviation of Differences = 3.16

p-value = 1.0

Again, this does not suggest that there is an agreement between instruments A and B. The inappropriate application of the test of significance, as a test for agreement, has also been discussed earlier in the article by Altman and Bland (1983). What matters in agreement is that each reading from the standard instrument should be repeated by the second instrument. We are not interested in the mean of readings by each instrument, but are interested in each individual reading. Therefore, comparing means using a significance test is not an appropriate method for assessing agreement.

2.2.2.4 Intra-class correlation coefficient

The Intra-class Correlation Coefficient or ICC, was devised initially to assess the relationship between variables within classes, or reliability. However, it was then used to assess agreement, to avoid the problem of linear relationship being mistaken for agreement in product moment correlation coefficient (*r*) (Bland & Altman, 1990; Lee et al., 1989). Different assignments of measurements of X and Y, in the calculation of the correlation coefficient (*r*), would produce different values of *r*. To overcome some of the limitations of the correlation coefficient (*r*), the ICC averages the correlations among all the possible ordering of the pairs (Bland & Altman, 1990). The ICC also extends to more than two observations, in contrast with the correlation coefficient (*r*) (Fay, 2005). In general, the ICC is a ratio of two variances:

$$\text{ICC} = \frac{\text{Variance owing to rated subjects}}{\text{Variance owing to subjects} + \text{Error}}$$

The value of the ICC can theoretically vary from 0 to 1, where 0 indicates no reliability or disagreement in the agreement study. The ICC of 1 indicates perfect reliability, or perfect agreement. There are different types of ICC that have been described by Shrout and Fleisse (1979). McGraw and Wong (1996) expanded the Shrout and Fleiss

system to include two more general forms of ICC. Weir (2005) summarised different types of ICC, based on models introduced by Shrout and Fleiss (1979), and McGraw and Wong (1996) (see Table 2.8).

Table 2.8: Different types of ICC

Shrout and Fleiss (1979)	**Computational formula**	**McGraw and Wong (1996)**	**Model**
1,1	$\frac{MS_B - MS_W}{MS_{B+}(k-1)MS_W}$	1	1-way random
1,k	$\frac{MS_B - MS_W}{MS_B}$	K	1-way random
	Use 3,1	C,1	2-way random
	Use 3,k	C, k	2-way random
2,1	$\frac{MS_S - MS_E}{(MS_S+(k-1)MS_E)+(k(MS_T-MS_E)/n)}$	A,1	2-way random
2,k	$\frac{MS_S - MS_E}{(MS_S+(k(MS_T-MS_E)/n)}$	A, k	2-way random
3,1	$\frac{MS_S - MS_E}{MS_S+(k-1)MS_E}$	C,1	2-way fixed
3,k	$\frac{MS_S - MS_E}{MS_S}$	C, k	2-way fixed
	Use 2,1	A,1	2-way fixed
	Use 2,k	A, k	2-way fixed

MS_B = between-subjects mean square; MS_E = error mean square; MS_S = subjects mean square; MS_T = trials mean square; MS_W = within-subjects mean square.

Shrout and Fleiss suggested three main models: model 1 is a one-way fixed model; model 2 is a two-way random model; and model 3 is a two-way fixed model (Shrout & Fleiss, 1979). The model is represented in the format of ICC (a,b). The value

of “a” can be 1, 2 or 3 (this depends on the three main models). For value “b”, when b=1, this suggests Single Measures ICC, and b=k suggests Averaged Measures ICC (Weir, 2005).

In the ICC model suggested by McGraw and Wong (1996), the designation “C” refers to consistency and “A” refers to absolute agreement. The “A” model considers both fixed and systematic error, whereas the “C” model only considers fixed error (McGraw & Wong, 1996; Weir, 2005).

Although a total of ten ICC models were summarised by Weir (2005) there are similarities in some of the ICC formula for different types of ICC (Weir, 2005). This is because the difference between the random model and the fixed model is not in the calculation, but in the interpretation of the ICC (McGraw & Wong, 1996).

According to Shrout and Fleisse (1979), there is only one ICC that measures the extent of absolute agreement, and that is ICC (2,1), which is based on the two-way random-effects ANOVA (Analysis of Variances) (Bruton et al., 2000; Shrout & Fleiss, 1979). This model is similar to ICC (A,1), as suggested by McGraw and Wong (1996) (Weir, 2005).

The ICC (C,1) for consistency simply compares the consistency between trials. For example, for the hypothetical data from Table 2.9, will produce ICC (C, 1) = 1.0, which is interpreted as a perfect agreement. However, the Absolute Agreement ICC, or ICC (A,1), compares both the consistency between trails and the agreement between ratings. So, the same pairs of data from Table 2.9 will produce ICC (A, 1) = 0.67, which suggests some degree of disagreement (or moderate agreement).

Table 2.9: Hypothetical dataset of repeated measurements from instrument A

Patient	**1st reading**	**2nd reading**
1	2	4
2	4	6
3	6	8

However, the use of ICC in assessing agreement has been criticised by Bland and Altman (1990). In testing the agreement of instruments, the new method will usually be compared to the standard instrument (Bland & Altman, 1990). The aim of testing is to ensure that the new method will produce the same measurements as the standard instrument (i.e. good agreement). This can also mean that the new method is designed to provide similar predictions of measurement as the standard instrument. So, there is clear ordering of the two variables, where the measurements from the standard instrument are usually denoted as X and measurements of the new method are denoted as Y.

The ICC ignores the ordering and treats both methods as a random sample from a population of methods (Bland & Altman, 1990). In an agreement study, there are two specific methods that will be compared, not two instruments chosen at random from some population. Another assumption in the ICC model, which is quite unjustified in methods comparison study, is that the measurement error of both methods has to be the same (Bland & Altman, 1990). The main purpose in testing agreement is to identify the measurement error of the new instrument in comparison to the standard instrument. Another issue with ICC is that it is influenced by the range of data. If the variance between subjects is high, the reliability will certainly appear to be high (Bruton et al., 2000).

2.2.2.5 Comparing slopes and y-intercepts

Often, in testing for agreement, the slope is tested against one. The argument is that if the two methods or instrument are equivalent (i.e. if it measures the same variable of the same subject, both instruments will give the same reading), thus the slope of the straight line will be one (Altman & Bland, 1983).

Straight line equation will show the relationship between two variables, and can be expressed as: $y = \alpha + \beta x$, where "y" is the predicted or expected value for any given value of "x", while "α" is the intercept of the straight line with the y-axis, and "β" is the slope. The values of both "α" and "β" are constant. The slope "β" is also called the *regression coefficient*, and measures the amount of change in the "y" variable for a unit change in "x" (Fay, 2005).

So, if instrument A measures "y", and instrument B measures "x", and if y=x, the slope of the straight line equation is equal to one. It is true that the straight line of y=x will always have slope of 1. However, this is not always true in reverse, because for a line with a slope of 1, the straight line could be y=x, or could be $y= \alpha+x$. Therefore, testing the slope is equal to 1, is also an inappropriate method of testing agreement.

When the test of slope is equal to 1 is significant, some people proceed to test the y-intercept. Theoretically, if slope is 1 and y-intercept is 0, then y will be equal to x (y=x). However testing both slope and intercept to assess agreement is not so popular compared to other methods.

Bland and Altman (2003) suggested another approach using the linear regression method in the evaluation of agreement. They suggested that the old measurement (y) can be regressed on the new measurement (x), and then one can calculate the standard error of a prediction of the old value from the new (Bland & Altman, 2003). This can be used to estimate predicted value from old measurements for any observed value of new measurement, with a confidence interval, which is also known as a *prediction interval*

(Bland & Altman, 2003). If we use data from Table 2.4 as an example, we can plot the laboratory values against the glucometer values, and obtain the regression line and prediction interval (see Figure 2.5). However, as noted by Bland and Altman in their paper (Bland & Altman, 2003), the problem is that the prediction interval is not constant; it is smaller in the middle, and wider towards the extremes. This is especially obvious for small samples in comparison with larger sample size (Bland & Altman, 2003).

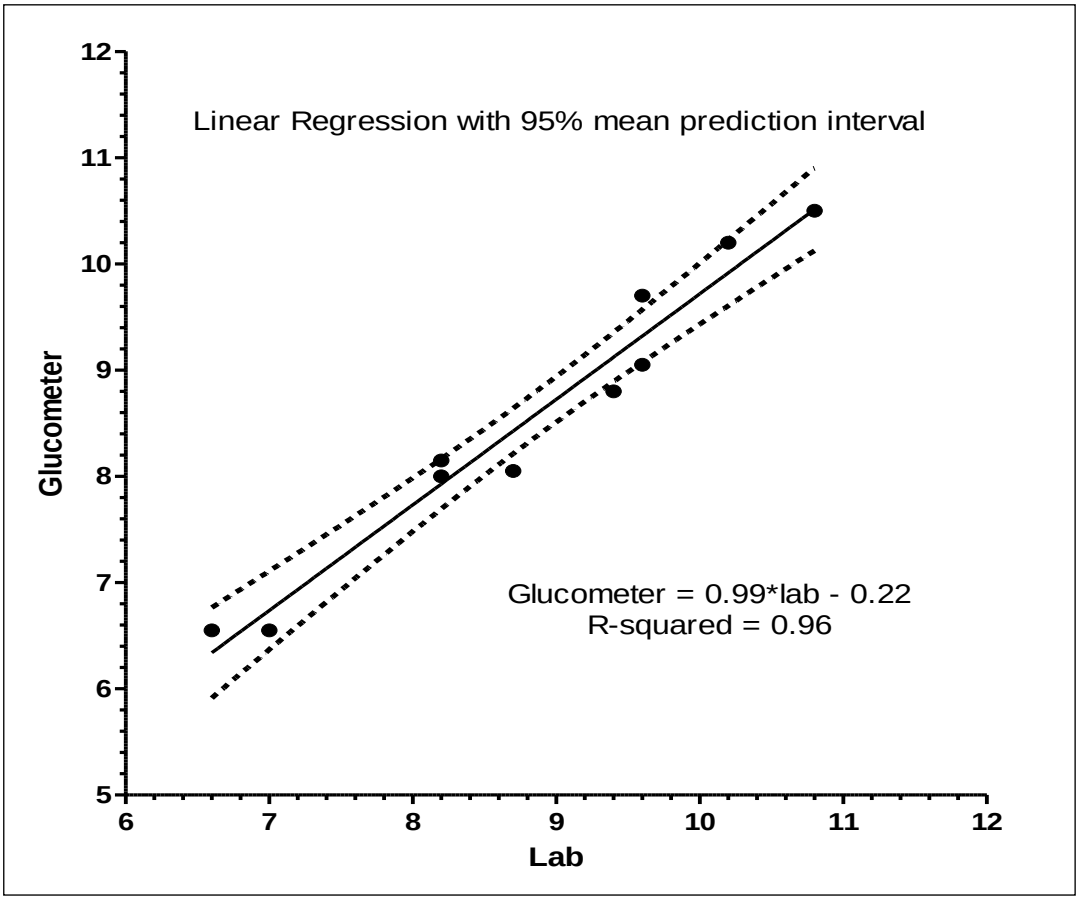


Figure 2.5: Linear Regression with Prediction Interval

2.3 Methods of Measuring Reliability

2.3.1 Systematic Review of Methods Used to Assess Reliability

This is the first ever systematic review on the statistical methods used to measure reliability of equipment measuring continuous variables in the medical literature. The aim of this review is to identify statistical methods used to assess the reliability of medical instruments measuring continuous variables in the medical literature. The proportion of various statistical methods found in this review will also reflect the level of knowledge (as determined by the statistical tests used) on the analysis of reliability. This review also follows the standards as suggested in the PRISMA statement (Moher et al., 2009). The PRISMA checklist for this review is attached as Appendix J.

2.3.1.1. Literature search and study selection

A search for literature was performed, in May 2010, from the electronic databases (Medline [EBSCOhost], Ovid, PubMed, Scopus and Science Direct) for studies investigating the reliability of instruments or equipment in medicine, published in journals between January 2007 and December 2009. Only full text articles were included in this review, and unpublished articles were not considered. Only studies that investigated the reliability of equipment measuring continuous variables were included.

The search term used was: Reliability AND (validation OR “comparison study”) AND medicine. The search also was limited to the medical area (including dentistry), studies involving human subjects, and articles written in English. Table 2.10 presents the summary of the literature search. All citations and abstracts were exported to the Endnote software, and then a search for duplicates was performed. Any studies with qualitative or categorical data, studies comparing instruments of different units, and association studies were excluded.

Table 2.10: Search of literature for reliability study

Database	Search criteria	Total hits	No. of related titles	No. of full articles available
Ovid	• Jan 2007–Dec 2009 • English • Human • Full text	334	35	23
Scopus	• Year 2007–2009 • English • Medicine (subject area) • Article (document type)	3,670	133	38
Medline (EBSCOhost)	• Jan 2007–Dec 2009 • Full text • Human • English	59	19	17
Science Direct	• Year 2007–2009 • Medical & Dentistry (subject area) • Journal article (document type)	1,599	86	84
PubMed	• 2007–2009 • English • Human • Full text	133	9	8
	TOTAL	**5,795**	**282**	**170**

2.3.1.2 Data extraction and analysis

Information on which statistical methods were used to assess reliability was extracted from each study. The statistical methods used were determined according to the information stated in the method section or the statistical analysis section, and also by identifying which statistical methods influenced the author's conclusion on the reliability of an instrument. Information on the year of publication and journal types was also extracted from each article. The journal types were divided into five areas: medicine (including emergency and critical care medicine); surgery; radiology; nutrition and others.

Descriptive analysis of the characteristic of studies and statistical methods used was performed. Univariate associations between statistical methods used and covariates (journal type, year of publication, and online databases) were assessed using Chi-square test and Fisher's exact test (where appropriate) with *p*-value of <0.05 was considered to be significant. All analyses were performed using SPSS 17.0 software.

2.3.1.4 Findings from the systematic review of reliability studies

A total of 5,795 titles were initially identified. However, after filtering for duplicates 5,563 titles and abstracts were reviewed. Only 282 were potentially related articles. A total of 170 full-text articles were reviewed. Of these, 131 articles did not meet the inclusion criteria, and a total of 42 articles were finally included in this review. Figure 2.6 summarises the selection process.

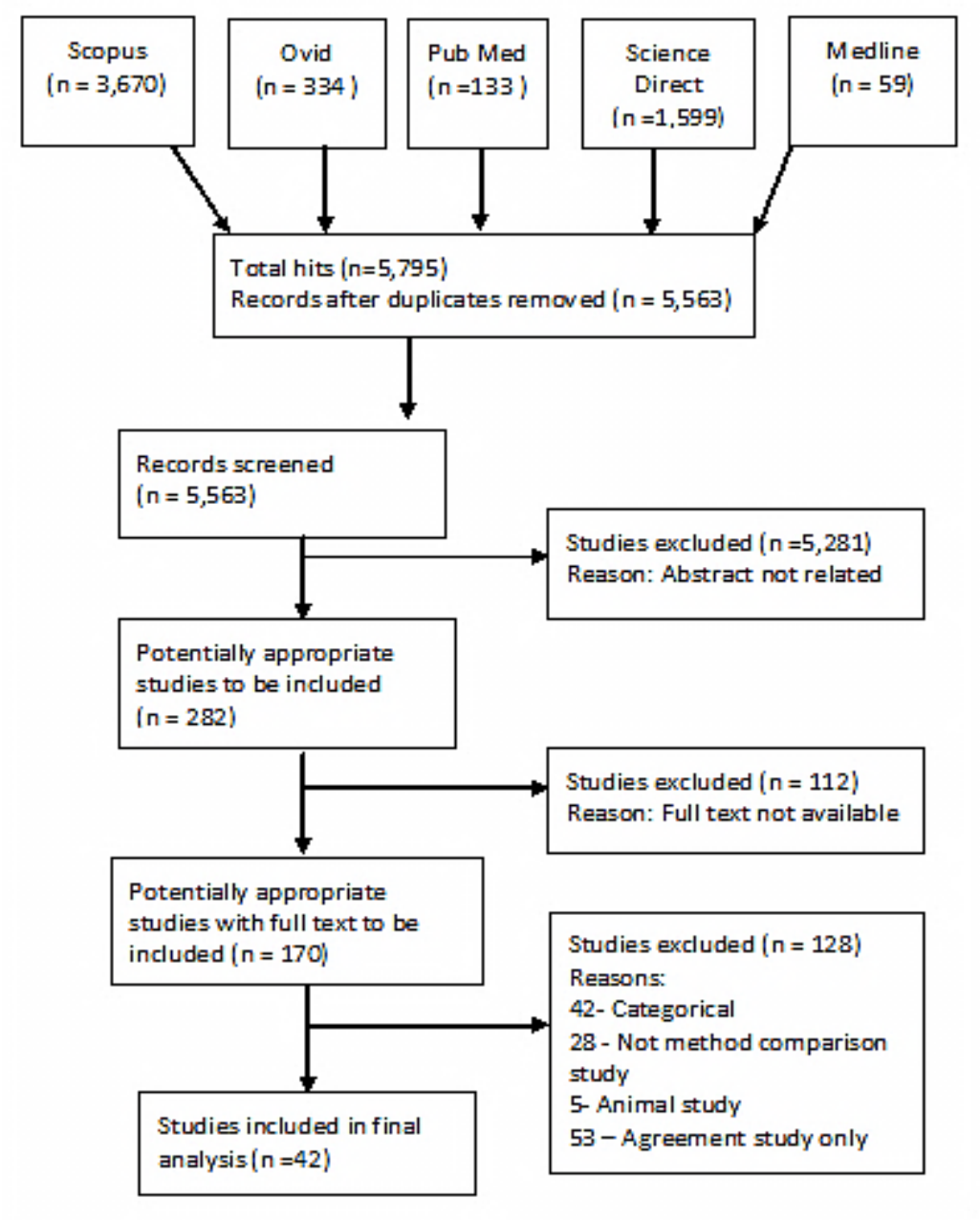


Figure 2.6: Selection of articles in reliability study

Out of the 42 articles reviewed, 26 (62%) were published in 2007, 7 (17%) in 2008, and 9 (21%) in 2009. Twelve (29%) articles were obtained from the Scopus database, 12 (29%) from the Science Direct database, 7 (17%) from the PubMed database, 6 (14%) from Ovid and 5(12%) from Medline. Most of the studies (32 or 76%) were published in medical journals, 4 (10%) in surgical journals, 3 (7%) in radiological journals, and 3 (7%) in dental journals.

Most of the reviewed studies (36 or 86%) relied on a single method to assess reliability. Others have used a combination of two or more methods. The Intra-class Correlation Coefficient was the most popular method used to assess reliability and was used in 25 (60%) of the reviewed studies. This was followed by the comparing means (8 or 19%), Bland-Altman Limits of Agreement (7 or 17%), and correlation coefficient (2 or 5%). These findings are shown in Table 2.11. Thirty studies (71%) also measured agreement at the same time. Out of 25 studies using the ICC, only 7 (28%) studies reported the confidence intervals and types of ICC used.

Table 2.11: Most popular statistical methods used to assess reliability in medicine

Statistical Method Used	**Number of methods used according to year of publication**			
	2007	2008	2009	Total
1. Intra-class Correlation Coefficient	6	5	14	25
2. Compare mean/ Mean difference	5	2	1	8
3. Bland Altman method (limits of agreement)	5	0	2	7
4. Correlation coefficient (*r*)	2	0	0	2

There was significant association between the year of publication and statistical method used, $p = 0.031$ (Fisher's exact). The use of Bland-Altman (limits of agreement) in assessing reliability reduced from five in 2007 to two in 2009. The use of ICC is becoming more popular, increasing from six in 2007 to 14 in 2009. The use of correlation coefficient (*r*) to assess reliability was only present in 2007. This suggests that researchers are aware that this method is not appropriate for assessing reliability.

Out of seven studies that used the Bland-Altman Limits of Agreement, four studies (Antona, Barra, Barrio, Gonzalez, & Sanchez, 2007; Boyles, Edwards, Gregory, Denman, & Clark, 2007; Holzinger et al., 2009; Shannon, Gregson, Stocks, Cole, & Main, 2009) used only the Band-Altman Limits of Agreement to measure reliability, two studies (Ageberg, Flenhagen, & Ljung, 2007; Maksymowych et al., 2007) used a combination with the ICC, and one study (Reilly K et al., 2007) used a combination with the correlation coefficient (*r*).

Total of two studies used correlation coefficient (*r*) to determine reliability found in this review. Out of this two, one study (Reilly K et al., 2007) used a combination of correlation coefficient (*r*) and the Bland-Altman Limits of Agreement, and another study (Syed, Oza, Vanderby, Heiderscheit, & Anderson, 2007) used only the correlation coefficient to conclude on the reliability of tested instrument. Two (5%) studies (Pini et al., 2008; Pini, Pastori, Baccheschi, Omboni, & Parati, 2007) have measured reliability using only the standard deviation of mean difference.

2.3.2 Review of Most Commonly Used Methods to Assess Reliability

2.3.2.1 Intra-class Correlation Coefficient

The Intra-class Correlation Coefficient was originally proposed by Sir Ronald Aylmer Fisher (R. A. Fisher, 1925). He was a statistician from England, and Fisher's exact test was one of his well-known contributions to statistics (J. Fisher, 1978). The earliest ICCs were modifications of the Pearson Correlation Coefficient (Weir, 2005). However, the modern version of ICC is now calculated using variance estimates, obtained from the analysis of variance or ANOVA, through partitioning of the total variance between and within subject variance (Bruton et al., 2000).

The general formula for ICC is given as (Weir, 2005):

$$\text{ICC} = \frac{\text{Subject variability } (\delta_S^2)}{\text{Subject variability } (\delta_S^2) + \text{Measurement error } (\delta_E^2)}$$

Values obtained from ANOVA table:

$$\text{Measurement error, } \delta_E^2 = \text{Mean square of Error, } MS_E$$

$$\text{Subject variability, } \delta_S^2 = \frac{\text{Mean square of Subject, } MS_S - \text{ Mean square of Error, } MS_E}{\text{Number of observer}}$$

The ICC is the most popular method used to assess reliability of medical instruments. There is no ordering of the repeated measures and can be applied to more than two repeated measurements (Streiner & Norman, 2003). As described in Section 2.2.2.4, ICC is a ratio of variances derived from ANOVA, so it is unit-less. The closer this ratio is to 1.0, the higher the reliability (Weir, 2005).

Suppose, for example, that we measure carbon monoxide level for ten patients using a same instrument three times. The hypothetical data are shown in Table 2.12. From the data, an ANOVA table can then be developed as in Table 2.13.

Table 2.12**:** Hypothetical data of repeated measurements of carbon monoxide level

Patient	**1st reading**	**2nd reading**	**3rd reading**	**mean**
1	6	7	8	7
2	4	5	6	5
3	2	2	2	2
4	3	4	5	4
5	5	4	6	5
6	8	9	10	9
7	5	7	9	7
8	6	7	8	7
9	4	6	8	6
10	7	9	8	8
mean	5	6	7	6

Table 2.13: Analysis of variance summary table

Source of variation	**Sum of squares**	**Degree of freedom**	**Mean square**
Patients	114	9	12.67
Raters/Instrument	20	2	10
Error	10	18	0.56
Total	144	29	

From the Table 2.13 the value of ICC can be calculated:

Measurement error, $\delta_E^2 = MS_E = 0.56$

Subject variability, $\delta_S^2 = \frac{MS_S - MS_E}{\text{Number of observer}}$

$$= \frac{12.67 - 0.56}{3} = 4.04$$

$$ICC = \frac{4.04}{4.04 + 0.56} = 0.88$$

The interpretation is that 88 per cent of the variance in the measurements results from the "true" variance among patients. However, note that this is according to the "classical" definition of reliability. There are different forms of ICC depending on various assumptions or criteria as described in Section 2.2.2.4.

Chinn (1991) recommended that any measure should have an Intra-class Correlation Coefficient of at least 0.6 to be useful Chinn, 1991). Rosner (Rosner, 2006) suggested the interpretation of ICC as shown in Table 2.14:

Table 2.14: Interpretation of ICC

ICC value	Interpretation
< 0.4	poor reliability
0.4 ≤ ICC < 0.75	fair to good reliability
≥ 0.75	excellent reliability

2.3.2.2 Comparing means/mean difference

Second most popular method that has been used to assess reliability is to compare means of two sets of measurements (either using t-test or looking at the mean difference). Since reliability involves repeated measurement of the same subject, a paired t-test is usually applied. However, the paired t-test only gives information about differences between the means of two sets of data, and not about individual differences (Bruton et al., 2000). As in the explanation in Section 2.2.2.3, on assessing agreement, comparing means is also not a suitable method of assessing reliability. Bruton et al. (Bruton et al., 2000) suggested that this test is not to be used in isolation, but may be complemented by other methods such as the Bland-Altman agreement analysis.

2.3.2.3 Bland-Altman Method

The Bland-Altman Limits of Agreement (LoA) also has been used as a method to assess reliability. Bland and Altman (1986) suggested that LoA are suitable for the analysis of repeatability of a single measurement method. However, the use of LoA to evaluate reliability has been criticised, as it only estimates reliability when there are two observations for each subject (Bland & Altman, 1986). This breaches the concept of reliability, that allows repeated (more than two) numbers of observations per subject (Fay, 2005). Although Bland and Altman (1999) suggested methods to deal with multiple measurements in calculating the LoA, this method is more suitable for the analysis of agreement rather than reliability. They proposed calculating the mean of the replicated measurements by each instrument, for each subject (Bland & Altman, 1999). Then, these pairs of means could be used to compare the two instruments using the limits of agreement (Bland & Altman, 1999).

The use of LoA in the analysis of reliability also has been criticised by Hopkins (2000), who gave reasons why LoA is not the best method to use for reliability analysis (Hopkins, 2000). According to Hopkins (2000), the values of the LoA can result in up to a 21% bias, and this depends on the degrees of freedom of the reliability study (i.e. number of participants and trials). Furthermore, Hopkins (2000) added that LoA cannot be applied to the simplest situation of only one trial (e.g. a urine test for a banned substance in an athlete).

2.3.2.4 Correlation Coefficient

As discussed in Section 2.2.2.2, the correlation coefficient provides information about the association and the strength of linear relationship. Correlation will not detect any systematic or fixed errors, and it is possible to have two sets of scores that are highly correlated, but not repeatable (Bruton et al., 2000). Therefore, it is recommended that the correlation coefficient should not be used in isolation for measuring reliability (Bruton et al., 2000; Neveu, Aubas, Seguret, Kramar, & Dujols, 2006). Furthermore, the correlation coefficient also breaches the concept of reliability, as it only estimates reliability when there are only two observations for each subject (Fay, 2005).

2.4 Issues in Method Comparison Studies

2.4.1 Agreement or Reliability?

Agreement and reliability are both important in assessing the quality of instruments. An instrument with high agreement will not be useful if it is unreliable. Ideally, these parameters should be assessed together. However, earlier systematic review showed that this is not commonly followed in practice, especially with respect to agreement studies. Most of the reliability studies (71%), found in the systematic review of earlier reliability studies also measured agreement at the same time. However, only 30% of agreement studies found in the systematic review of the agreement studies assessed reliability. Researchers tend to focus on one aspect of quality when validating instruments, although there is a possibility of agreement and reliability studies being conducted separately for the same instrument. Nonetheless, it is important to ensure the reliability of the instrument first, before testing for agreement, because it is impossible to assess the agreement of an unreliable instrument.

2.4.2 Single or Multiple methods?

According to both systematic reviews conducted earlier in this chapter, most reliability studies (86%) relied on a single statistical method to assess reliability, in contrast with agreement studies where most of the studies (65%) used a combination of statistical methods (see Table 2.15). A strong case for using multiple methods in assessing agreement and reliability is because each statistical method has its own strengths and weaknesses. The usage of multiple methods has the advantage of compensating for the limitations of any one single method. As long as the methods chosen are appropriate for it purposes. Luiz and Szklo (2005) suggested that more than one statistical method to assess agreement may be reported usefully, since no strategy seems to be fool proof (Luiz & Szklo, 2005). Similarly, in reliability studies, it was suggested that no single

reliability estimate should be used for reliability studies, and a combination of methods was more likely to provide more information on the reliability of an instrument (Bruton et al., 2000).

However, another possible reason for using multiple methods is the researcher's limited understanding of the statistical methods for agreement and reliability. This is probably the reason for the application of multiple inappropriate statistical methods in a single study; for example, the use of both correlation coefficient and significance test of the difference between means, to test for agreement and reliability. Both of these methods have been clearly shown to be inappropriate statistical methods to assess agreement and reliability (Altman & Bland, 1983; Daly & Bourke, 2000).

Table 2.15: Single versus multiple methods

	AGREEMENT (N=210)	**RELIABILITY (N=42)**
Overall:		
• Multiple methods	137 (65%)	6 (14%)
• Single method	73 (35%)	36 (86%)
$p<0.0001$		
According to year:		
2007	n=70	n=26
• Multiple methods	43 (61%)	6 (23%)
• Single method	27 (39%)	20(77%)
$p=0.0002$		
2008	n=70	n=7
• Multiple methods	46 (66%)	0
• Single method	24 (34%)	7 (100%)
$p=0.0009^{*}$		
2009	n=70	n=9
• Multiple methods	48 (69%)	0
• Single method	22 (31%)	9 (100%)
$p<0.0001^{*}$		
(*Fisher's exact)		

2.4.3 Application of Inappropriate Statistical Methods

The proportion of studies with inappropriate statistical methods, found in both earlier systematic reviews, will reflect the proportion of medical instruments that have been validated using inappropriate methods in current clinical practice. As found in the earlier systematic reviews, eight (19%) of reliability studies and twenty (10%) of agreement studies used inappropriate methods, which means that there is a distinct possibility that some medical instruments or equipment used currently were validated using inappropriate methods, with consequently erroneous conclusions being drawn from these methods. This equipment, therefore, may not be as precise or accurate as believed, which could, potentially, affect the management of patients, the quality of care given to patients and, worse, it could cost lives. Inappropriate application of statistical methods in method comparison studies also reflects the lack of knowledge in this area among medical researchers. This is alarming and it is important for clinicians or medical researchers to be aware of this.

2.4.4 Is the most popular method the best?

2.4.4.1 Agreement Analysis

Although the Bland-Altman Limits of Agreement is the most popular method used to assess agreement, there are a few issues and limitation related to it of which medical researchers should be aware of.

a. Confidence Interval for Limits of Agreement

Limits of agreement is actually just an estimate of the values which apply to the whole population (Bland & Altman, 1987). So, whatever value of limits of agreement are obtained from a study, they only apply to that study population. If a similar study was repeated in a different study population, this second sample would give different limits of agreement. Therefore, to infer the limits of agreement to the whole population, a 95% confidence interval (CI) of the upper and lower limit of agreement should be calculated, as suggested by Bland and Altman (Bland & Altman, 1987). The 95% confidence intervals can be calculated by finding the appropriate point of the *t* distribution with *n* - 1 degrees of freedom and the standard deviation of the difference, SD (Bland & Altman, 1987):

$$\text{CI for upper limit of agreement} = \text{Mean Bias} + (1.96(\text{SD}) \pm t\sqrt{\frac{3\text{SD}^2}{\text{n}}});$$

$$\text{CI for lower limit of agreement} = \text{Mean Bias} - (1.96(\text{SD}) \pm t\sqrt{\frac{3\text{SD}^2}{\text{n}}});$$

However, this is rarely practised by researchers. Out of 178 papers reviewed earlier that used the Bland-Altman method to assess agreement, only one paper considered the 95% confidence interval of limits of agreement. Bland and Altman are also aware of this problem and regret that these confidence intervals are seldom quoted (Bland & Altman,

2003). Theoretically, without reporting the confidence interval, their conclusion about the agreement of methods measured can only be applied to the measurement during the research, and cannot be inferred to clinical practice.

This issue has also been discussed in detail by Hamilton and Stamey (2007), who suggested that Limits of Agreement only provide a reference interval, and can be misleading if the Confidence Interval (CI) is not considered (Hamilton & Stamey, 2007). They concluded that Limits of Agreement should never be used as the decisive factor in concluding agreement between two instruments (Hamilton & Stamey, 2007).

b. Interpretation of Bland-Altman Limits of Agreement

One of the reasons why the Bland-Altman Method is so popular is its simplicity (M. D. Cohen & Jennings, 2002). Although the interpretation of limits of agreement seems to be simple and easy, medical researcher should be aware of the appropriate way of interpreting the Bland-Altman analysis. Mistakes or inappropriate interpretation of limits of agreement can occur as found in the following published article.

In 2005, a study tested the agreement of three peak flow meters (A, B and C) using three statistical methods (Pearson's Correlation Coefficient, t-test, and the Bland-Altman method) (Nazir et al., 2005). For peak flow meters A and B, the limits of agreement were found to be 40 l/min to 60 l/min. The authors interpreted this as the differences between peak flow meter A and B to range from 40–60 l/min (Nazir et al., 2005). They did not comment whether peak flow B would overestimate the value of peak flow A, which is the most important clinical finding desired. Furthermore, the overall conclusions on the agreement of the peak flow meters were made based on a paired t-test.

In fact Bland and Altman themselves made a mistake in the interpretation of the limits of agreement in one of their earlier publications (Bland & Altman, 1987), where

they compared the readings between a large peak flow meter (PEFR) and mini peak flow meter. By plotting the difference (Large PEFR – mini PEFR) against the mean, the upper limit of agreement was 75.5 l/min and the lower limit of agreement was -79.7 l/min (Bland & Altman, 1987). Their interpretation was that the mini peak flow meter may be 80.0 l/min below or 76.0 l/min above the large peak flow meter. However, because the difference was calculated from Large PEFR – mini PEFR, the positive difference means that the mini PEFR underestimates the large PEFR, and the negative difference means that the mini PEFR overestimates the large PEFR. So, the appropriate interpretation should be that the mini PEFR may be 80.0 l/min above or 76.0 l/min below the large PEFR.

Thus, a mix of negative and positive values of limits of agreement might confuse some researchers. In addition, imagine if we apply the 95% confidence interval for the limits of agreement. This would create further confusion and make the Bland-Altman method appear to not be as straightforward as originally thought. Therefore, medical researcher should put an effort to really understand this method and interpret the result appropriately.

c. Proportional Bias

Hopkins (2004) demonstrated that the Bland-Altman plot indicates incorrectly that there is a systematic bias in the relationship between two measures (Hopkins, 2004). Using a fixedly generated data, Hopkins clearly showed the proportional bias produced in the Bland-Altman plot, but not in the regression (ordinary least squares method) analysis. If a slope of regression line fitted to the Bland-Altman plot differs significantly from zero, it is argued that proportional bias exists (Ludbrook, 2002). Using randomly generated data, Hopkins showed that proportional bias was produced in the Bland-Altman plot, but not in the regression (ordinary least squares method) analysis, and concluded that

the Bland-Altman plot should not be used to make conclusions about bias for any instrument (Hopkins, 2004). He added that bias in the Bland-Altman plots was not restricted to calibrated instruments, but could arise as an artefact of random error between measures that have not been calibrated (Hopkins, 2004). Commenting on Hopkins' article, Batterham (2004) favoured the ordinary least squares regression technique, rather than the Bland-Altman limits of agreement (Batterham, 2004).

However, Ludbrook (2002) claimed that the presence of bias in the analysis was a result of some kind of statistical assumption, and suggested that an approach using least-products regression to fit the regression line in the Bland-Altman plot apparently eliminated the bias problem in Bland-Altman plots (Ludbrook, 2002).

The main concern about the proportional bias is that this will result in artefactual bias in the prediction. The predicted bias will consist of artefact and real bias, which cannot be differentiated by the researcher (Hopkins, 2004). It is recommended that a linear regression line should be fitted to the Bland-Altman plot, and the use of ordinary least squares regression analysis to test for the proportional bias is accepted (Ludbrook, 2002). If the slopes of the line are not significantly different from zero then the proportional bias is absent.

2.4.4.2 Reliability Analysis

Intra-class Correlation Coefficient or ICC is the most popular method used to assess the reliability of medical instruments. There are a few concerns regarding the application of ICC in evaluating reliability:

a. Choosing appropriate type of ICC

There are different types of ICC, and confusion exists regarding which ICC to use (Weir, 2005). Muller and Buttner (2004) demonstrated that different types of ICC may

result in quite different values for the same dataset, under the same sampling theory (Muller & Buttner, 1994). So it is important to determine which type of ICC is suitable, depending on the purpose of the analysis. Weir (2005) suggested some issues that should be considered when choosing an ICC test:

(a) One- or two-way model:

- For the one-way model each subject is assumed to be assessed by different raters, and the raters are also assumed to be selected from the population. This model allows for situations where all subjects are not rated by all raters. In this model, all sources of error are lumped together. A one-way model should be considered when information on which raters rated the subject is not known (Weir, 2005).
- The two-way model assumes that each subject was assessed by the same raters, and requires raters to be crossed with subjects (i.e. each rater rates all subjects). The two-way model allows the error to be devised into random and fixed errors (Shoukri & Pause, 1999; Weir, 2005).

(b) Random- or fixed-effect model

- In a fixed-effects model, the levels of variable are fixed or specified in advance (Rosner, 2006). The fixed factor is considered when all levels of the factor of interest are included in the analysis. Raters are considered as fixed effects, but items/subjects are treated as random effects (no generalization beyond the sample). So, there is no attempt to generalise the result on reliability (Weir, 2005).
- Under a random-effects model, both factors (raters and items/subjects) are viewed as random effects (Rosner, 2006). Random factor is considered when the analysis is to be generalised to other levels (Weir, 2005).

(c) Single or mean score (Weir, 2005):

- Single Measures ICC should be reported if only a single measure on a subject was taken.
- If two or more trials were measured on a subject, then Average Measures ICC should be reported. The Averaged Measures ICC will always be higher than the Single Measures ICC

b. Between-subjects variability

The ICC is influenced greatly by between-subjects variability. If the ICC is applied to data from a group of individuals with a wide range of the measured characteristics, the value of the ICC will indicate higher reliability, compared to the same analysis when applied to a group of data with a narrow range of the same characteristic (Weir, 2005). However, according to Weir (2005) this is an unfair criticism, because the ICC is not meant to provide an index of absolute measurement error (Weir, 2005). In general, the ICC is a ratio and does not quantify precision.

2.5 Proposed Method of Measuring Agreement

Simplicity, practicality (or interpretability), and ability of a certain method to detect systematic bias are among the important factors when choosing a method to evaluate agreement (Luiz & Szklo, 2005). While detecting bias has been the main focus, simplicity and practicality are also important, because the analysis will also be used and interpreted by non-statistical audiences (for example medical researchers and clinicians). This has contributed to the popularity of the Bland-Altman method (Luiz & Szklo, 2005). However, with the limitation of Bland-Altman method (as discussed in Section 2.4.4.1), there is a need for an alternative approach in evaluating agreement.

2.5.1 Comparison of slopes and y-intercepts

In this study, a method of measuring agreement, based on the comparison of slopes and y-intercept, will be explored. This consists of a comparison of a linear regression line ($y=\alpha+\beta x$) with the line of agreement ($y=x$). A linear regression equation is used to predict the "y" value from the value of "x".

Regression is a method used for estimating the numerical relationship between variables (McPhillips-Tangum, Aubert, Bailey, & Koplan, 1997), and will give the "best-fit" line to the set of data points plotted in a scatter diagram. This line is called a *regression line* (Fay, 2005). The corresponding equation to the regression line is called a *regression equation* (Fay, 2005). The most popular method to fit a linear regression line is the Ordinary Least Square (OLS) method, first introduced by Adrien Marie Legendre, a French mathematician, in 1805 (Zar, 2010). Although the German mathematician Carl Friedrich Gauss claimed that he had used the method at least ten years before that, the term "least squares" is credited to Legendre's publication of 1805 (Zar, 2010).

The OLS method minimises the sum of squared vertical distances between the observed values in the dataset, and the predicted values by the linear approximation (Zar, 2010). The use of OLS regression for comparing methods of measurement has been criticised by Ludbrook (2000) who pointed out two main reasons why OLS is not appropriate. Firstly, the assumption of OLS regression that the values of the predicted "y" variable are attended by fixed error, and the values of the predictor "x" variable are fixed in advanced (i.e. without fixed error), was rarely met in method comparison studies (John. Ludbrook, 2002). Secondly, the values of "y" variable are regarded as the "gold standard" or the "benchmark". Ludbrook claimed that when two methods are compared in methods comparison study, neither method can be considered as the benchmark (John. Ludbrook, 2002). To solve these issues, Ludbrook suggested the use of Ordinary Least Product (OLP) instead (John. Ludbrook, 2002).

However, it seems that Ludbrook misunderstood the concept of method comparison studies. One of the main purposes of conducting the methods comparison study is to ensure that the new instrument or method is able to provide similar measurements or predictions as the standard instrument, or currently used instrument. So, the measurement from the standard instrument will definitely be the referral or standard measurement (the predictor "x"), and the value of "y" from the new or tested instrument is the predicted value, and dependent on the standard value "x". Hopkins (2004) and Batterham (2004) also favoured OLS rather than OLP regression for application in method comparison studies (Batterham, 2004; Hopkins, 2004). Therefore, theoretically the use of OLS regression is suitable to be applied in the analysis in method comparison study.

Perfect agreement occurs when two instruments measuring the same variable of the same subject, produce a similar result. If "y" is the value from a new instrument, and "x" is the value from the standard instrument, plotting the value of y against x will produce a line with linear equation of: $y = \alpha + \beta x$; where $\beta = \frac{\sum xy}{\sum x^2}$ and $\alpha = \frac{\sum y}{n} - \beta\left(\frac{\sum x}{n}\right)$ (Zar, 2010).

In the situation of perfect agreement, where $y = x$ then the line will have a slope $\beta = 1$, and intercept $\alpha = 0$. The straight line ($y = \alpha + \beta x$) can be compared with the line of agreement ($y = x$), and any differences in the intercept and slope can be tested. If there is no difference between the two lines, this means that the two instruments agree.

This is very similar to the concept that was suggested by Passing and Bablok (1983), who described a method based on linear regression procedure, with no special assumptions on the distribution of the samples and the measurement errors (Passing & Bablok, 1983). The Passing and Bablok method does not depend on the assignment of the methods to X and Y. They calculated the slope and intercept with 95% confidence intervals, and these confidence intervals were then used to test if there was any difference between slope and 1, and between intercept and 0. However, their method is not so popular.

To compare the linear regression lines, one first needs to compute the two straight lines $y_1 = \alpha_1 + \beta_1 x_1$ and $y_2 = \alpha_2 + \beta_2 x_2$. Then, the next step is to compare the slopes of the two regression lines, by testing the null hypothesis (H_0): the two slopes are identical ($\beta_1 = \beta_2 = 1$). If the *p*-value <0.05, then the lines are significantly different. So, there is no point in comparing the intercepts because the lines are clearly not the same. Thus, this suggests that there is no agreement between the two instruments or methods.

In the situation of *p*-value >0.05, this means that the slopes are not significantly different. If the slopes are indistinguishable, this suggests that the lines could be parallel with distinct intercepts, or the lines could be identical, with the same slopes and intercepts. So, the next step is to test the null hypothesis (H_0): the two intercepts are identical (i.e. the intercept = 0). If the *p*-value for this test is <0.05, this suggests that there is a significant difference in the intercept of the two lines. This means that the lines are not the same (they are distinct but parallel). If *p*-value is >0.05, then there is no evidence that can suggest that the lines are different.

2.5.1.1 Comparing Slopes

a. Student's t-test

A simple method of testing the null hypothesis about the equality of two slopes $H_0: \beta_1 = \beta_2$ involves the use of Student's t-test (Zar, 2010). In the case of comparing with the line of agreement $\beta_2 = 1$. The test statistics is:

$$t = \frac{\beta_1 - \beta_2}{S_{\beta_1 - \beta_2}}$$

where,

$S_{\beta_1 - \beta_2}$ = standard error of the difference between slopes

$$S_{\beta_1 - \beta_2} = \sqrt{\frac{(s^2_{Y \cdot X})_p}{(\sum x^2)_1} + \frac{(s^2_{Y \cdot X})_p}{(\sum x^2)_2}}$$

$s^2_{Y.X_p}$ = Pooled residual mean square

$$\text{Pooled residual mean square} = \frac{(\text{residual SS})_1 + (\text{residual SS})_2}{(\text{residual DF})_1 + (\text{residual DF})_2}$$

$$\text{Residual SS} = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2}$$

$$\text{DF} = \text{degree of freedom} = \text{n} - 2$$

The critical value of *t* for this test has $(n_1 - 2) + (n_2 - 2)$ degree of freedom, or the critical value $(v) = n_1 + n_2 - 4$.

If H_0 is not rejected, then the common regression coefficient can be calculated:

$$\text{Common slope, } \beta_c = \frac{(\sum xy)_1 + (\sum xy)_2}{(\sum x^2)_1 + (\sum x^2)_2}$$

To demonstrate this, an example of analysis from (Zar, 2010) is given below:

H_0: $\beta_1 = \beta_2$

H_1: $\beta_1 \neq \beta_2$

Group 1

n_1 = 26 $(\sum x^2)_1$=1470.8712 $(\sum y^2)_1$=13299.5296

$(\sum xy)_1$= 4363.1627

β_1= 4363.1627/1470.8712 = 2.97

Residual SS_1 = 13299.5296 – [$(4363.1627)^2$/1470.8712] = 356.7317

Residual DF_1 = 26 – 2 = 24

Group 2

n_2 = 30 $(\sum x^2)_2$=2272.4750 $(\sum y^2)_2$=10964.0947

$(\sum xy)_2$=4928.8100

β_2= 4928.81/2272.48= 2.17

Residual SS_2 = 10964.0947 – [$(4928.8100)^2$/2272.4750] = 273.9142

Residual DF_2 = 30 – 2 = 28

The pooled residual mean square, $s^2_{Y.X_p}$= (356.7317 + 273.9142)/(24 + 28) = 12.1278

Standard error of the difference between slopes,

$$S_{\beta_1-\beta_2} = \sqrt{(12.1278/1470.8712) + (12.1278/2272.4750)} = 0.1165$$

t = (2.97 – 2.17)/0.1165 = 6.867

ν = 24 + 28 = 52

Reject H_0 if $|t| \geq t_{\propto(2),\nu}$

$t_{0.05(2),52}$= 2.007; therefore reject H_0 (p-value < 0.001)

b. Analysis of covariance (ANCOVA)

Analysis of covariance (ANCOVA) can also be used to compare two lines where H_0: β1 = β2 and H_1: β1 ≠ β2 (T. P. Smith, 2012). This analysis also can be used if the slopes of more than two lines are to be compared (Zar, 2010). The basic caculation require quantities already computed: $\sum x^2$, $\sum xy$, $\sum y^2$ (i.e. total sums of squares, SS), and the residual SS and degree of freedom (DF) for each line, as shown in the Table 2.16 (Zar, 2010).

Table 2.16: Calculation for testing for significant differences among slopes

	$\sum x^2$	$\sum xy$	$\sum y^2$	**Residual SS**	**Residual DF**
Regression 1	A1	B1	C1	$SS_1 = C_1 - \frac{B_1^2}{A_1}$	$DF_1 = n_1 - 2$
Regression 2	A2	B2	C2	$SS_2 = C_2 - \frac{B_2^2}{A_2}$	$DF_2 = n_2 - 2$
.	.	.	.		
.	.	.	.		
.	.	.	.		
Regression k	A_k	B_k	C_k	$SS_k = C_k - \frac{B_k^2}{A_k}$	$DF_k = n_k - 2$
Pooled regression				$SSp = \sum_{i=1}^{k} SS_i$	$DF_p = \sum_{i=1}^{k} n - 2k$
Common regression	$A_C = \sum_{i=1}^{k} A_i$	$B_C = \sum_{i=1}^{k} B_i$	$C_C = \sum_{i=1}^{k} C_i$	$SS_C = C_C - \frac{B_C^2}{A_C}$	$DF_C = \sum_{i=1}^{k} n_i$ – k - 1
Total regression	A_t	B_t	C_t	$SS_t = C_t - \frac{B_t^2}{A_t}$	$DF_t = \sum_{i=1}^{k} n_i - 2$

F-test then can be used to test the null hypothesis (Zar, 2010):

$$F = \frac{\left(\frac{SS_C - SS_p}{k-1}\right)}{\frac{SS_p}{DF_p}}$$

This can be demonstrated using data from Table 2.17. Tabe 2.18 can be build based on formula descrbed by Zar (2010) and Lowry (2012) (Lowry, 2012; Zar, 2010) :

Table 2.17: Hypothetical data to demonstrate ANCOVA

	X	**Y_1**	**Y_2**
1	12.9	4.44	4.81
2	16.0	4.02	3.99
3	20.1	3.68	3.68
4	22.5	3.02	2.93
5	25.0	2.65	2.91
6	27.5	2.21	2.58
Mean	20.67	3.337	3.483
Sum	124	20.02	20.9

Table 2.18: ANCOVA table

	DF	**SSX2**	**SSXY**	**SSY2**	**Residual SS**
Regression 1	4	151.33	-23.2067	3.6399	0.0811
Regression 2	4	151.33	-22.4533	3.5083	0.1768
Pooled regression	8	-	-	-	0.2579
Common regression	9	302.66	-45.66	7.1482	0.2598
Total regression	10	302.66	-45.48	7.2128	0.3244

$$F = \frac{\left(\frac{SS_C - SS_p}{k-1}\right)}{\frac{SS_p}{DF_p}} = \frac{0.2598-0.2579}{\frac{0.2579}{8}} = 0.0589$$

As F=0.00589, DFn = 1, DFd = 8, $p > 0.50$, therefore do not reject H_0

2.5.1.2 Comparing intercepts

a. Student's t-test

To test the null hypothesis about the equality of two intercepts (**H_0: $\alpha_1=\alpha_2$**). In the case of comparing with the line of agreement, $\alpha_2 = 1$. The test statistic is (Zar, 2010):

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - \beta_c(\bar{X}_1 - \bar{X}_2)}{\sqrt{s^2_{X.Y_c}\left(\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{X}_1 - \bar{X}_2)^2}{A_C}\right)}}$$

where,

$\bar{Y}$ = mean Y

$\bar{X}$ = mean X

β_c= regression coefficient for common regression

$s^2_{X.Y_C}$ = residual MS for common regression $= \frac{SS_c}{DF_c}$

SS_c = residual SS for common regression $= C_c - \frac{B_c^2}{A_c}$

C_c = sum of squares of Y for common regression= $(\sum y^2)_1 + (\sum y^2)_2$

B_c = sum of crossproducts for common regression= $(\sum xy)_1 + (\sum xy)_2$

A_c = sum of squares of X for common regression= $(\sum x^2)_1 + (\sum x^2)_2$

Example of analysis (Zar, 2010):

Group 1

n=13 $\bar{X}$=54.65 $\bar{Y}$=170.23 $\sum x^2$= 1012.1923 $\sum xy$ =1585.3385

$\sum y^2$=2618.3077 β=1.57 α=84.6 residual SS=135.2833

residual DF=11

Group 2

n=15 $\bar{X}$=56.93 $\bar{Y}$=162.93 $\sum x^2$= 1659.4333 $\sum xy$ =2475.4333

$\sum y^2$=3848.9333 β=1.49 α=78.0 residual SS=156.2449

residual DF=13

Slope test:

H_0: $\beta_1 = \beta_2$

H_1: $\beta_1 \neq \beta_2$

t = 0.575, v = 24

$t_{0.05(2),24}$ = 2.064; therefore do not reject H_0 (p-value = 0.57)

Test for intercept:

H_0: $\alpha_1 = \alpha_2$

H_1: $\alpha_1 \neq \alpha_2$

A_c = 1012.1923 + 1659.4333 = 2671.6256

B_c = 1585.3385 + 2475.4333 = 4060.7718

C_c = 2618.3077 + 3848.9333 = 6467.2410

B_c = 4060.7718//2671.6256 = 1.520

SS_c = 6467.2410 – (4060.7718)^2/2671.6256 = 295.0185

DF_c = 13 + 15 - 3 = 25

$s^2_{X.Y\,C}$ = 295.0185/25 = 11.8007

$$t = \frac{(170.2 - 162.93) - 1.520(54.65 - 56.93)}{\sqrt{11.8007\left(\frac{1}{13} + \frac{1}{15} + \frac{(54.65-56.93)^2}{2671.6256}\right)}} = \frac{10.77}{1.3105} = 8.218$$

$t_{0.05(2),25}$ = 2.060; therefore reject H_0 (p-value < 0.001)

b. Analysis of covariance (ANCOVA)

ANCOVA also can be used for testing whether the elevations are equal for two regression lines. The null hypothesis is that the elevations are equal (i.e. the lines coincide) while the alternative hypothesis is that the elevations are not equal (i.e. the lines do not coincide) (T. P. Smith, 2012). The test statistic for testing the null hypothesis is given as (Zar, 2010):

$$F = \frac{\left(\frac{SS_t - SS_C}{k-1}\right)}{\frac{SS_C}{DF_C}}$$

Using the same information from Table 2.17 and Table 2.18,

$$F = \frac{0.3244 - 0.2598}{\frac{0.2598}{9}} = 2.2379$$

As F=2.2379, DFn = 1, DFd = 9, p > 0.50, therefore do not reject H_0

2.5.3 Agreement Model

The aim of the agreement study is to identify or predict the error of the new instrument. Conclusions from the comparison of slopes and y-intercepts analysis (i.e. equal slopes or equal intercepts) will not provide direct information on the magnitude of error or bias. So, an agreement model is proposed to quantify any error or bias produced.

The true value of a measurement, *x*, is estimated or predicted by *y* which differs from the true value by an error or bias (Hibbert, 2007). An error is viewed as having two components, namely, a random component and a systematic component (JCGM, 2008). Since *Predicted (y) = True value (x) + Error*, a function of error is given as;

$$Error = Prediceted\ value - True\ value$$

Predicted value is measurement obtained from the new instrument (y) and true value is measurement obtained from the standard instrument (x). So, $Error = y - x$.

Since $y = \alpha + \beta x$, thus $Error = y - x = \alpha + \beta x - x = \alpha + (\beta - 1)x$

Therefore error or bias can be estimated using this function. Details of the proposed analysis will be described in Chapter 3.

2.6 Summary of Chapter 2

This chapter presents the first systematic review that identifies the most common statistical methods used to assess agreement and reliability of equipment measuring continuous variables in recent studies (in medicine). There are several methods and approaches that have been used to measure agreement. The most common method to assess agreement is the Bland-Altman Limits of Agreement (LoA), followed by Correlation Coefficient (*r*), comparing means, comparing slope and intercept, and Intra-class Correlation Coefficient. Various methods have also been used to estimate reliability, and among these popular methods include: Intra-class Correlation Coefficient, comparing means, Bland-Altman Limits of Agreement, and Correlation Coefficient (*r*).

The statistical methods used to assess agreement and reliability, found in the review, can be used as a surrogate measure of statistical knowledge on method comparison studies among medical researchers. Moreover, the proportion of various statistical methods found in this review will reflect the proportion of medical instruments that have been validated, using those particular statistical methods in current clinical practice.

This chapter also reviews the theoretical aspects of the most commonly used methods, and points out that some of the methods are inappropriate to be used in method comparison study. Some of the methods that were found to be inappropriate in assessing agreement include the Correlation Coefficient (*r*), comparing means, and ICC. In the analysis of reliability, Correlation Coefficient (*r*), Bland-Altman Limits of Agreement and comparing means were thought to be inappropriate. In addition, there is no single method that is fool proof, and even the most popular method such as Bland-Altman Limits of Agreement, has been criticised for its weaknesses.

Section 2.4 highlighted a few issues related to the methods comparison study, which includes the evaluation of agreement and reliability in a single study, the application of multiple statistical methods, and the use of inappropriate methods in testing agreement and reliability. Since the methods used to assess agreement has been criticised the most, and no strong method that has been found for the evaluation of agreement, analysis of agreement based on linear regression method was explored, and an agreement model to quantify bias was proposed in Section 2.5.

Finally, findings from the two systematic reviews, and issues in the methods comparison study highlighted in this chapter, will be an important contribution to medical research. Although there is no single perfect method, researchers should be aware of the inappropriate methods that they should avoid when analysing data in method comparison studies (i.e. to assess agreement and reliability). This is important because inappropriate analysis will lead to invalid conclusions. However, further analysis is required to compare different statistical methods used in method comparison study, before any recommendation or definite conclusion can be made.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This chapter describes the research methodology used in this study. It depicts the flow of the study, starting with the methods used for data collection and the method used to generate the findings in order to achieve the objectives of the study. Section 3.2 explains the design and population of this study. This is followed by Section 3.3 which describes the study variables. Section 3.4 describes the study instrument and explains the measurement procedure. The processes of ethical clearance and application for funding are described in Section 3.5. Section 3.6 explains the estimation of sample size for this study and explains how the calculation was performed. Details on data collection are elaborated under Section 3.7. This chapter also describes all statistical software and the statistical methods used for data analysis in this research under Sections 3.8 and 3.9. Data management and data analysis in this study are described in Section 3.10 and Section 3.11.

3.2 Study Design and Study Population

A cross-sectional study was conducted to assist with data collection for this study. Data were collected from two population settings:

- **Institutional setting (University of Malaya population)** – Participants of the UM Wellness Health-Screening Programme and UM Wellness Quit Smoking Clinic.
- **Community-based setting** – Participants of the community health-screening programme in Kampung Teluk Gadong Kecil, Klang, Selangor and community health screening at Mid Valley Megamall, Kuala Lumpur.

Data were collected from multiple centres to ensure variation in the range of data for all variables. Samples from the institutional setting are limited to the working-age group, and also limited to the criteria set by the wellness programme of the institution. Samples from the community setting will include a wider age group, including youngsters and the elderly. The elderly population has a different range of variables compared to the youngsters or the working-age group. For example, the values of blood pressure in elderly patients are more likely to be abnormal or higher than the normal population or the young population. This scenario is similar for most clinical variables.

A convenient sampling method was applied to all the participants attending the health-screening programmes. Only participants that were willing to participate in this study were included. Due to limited resources and time constraints, only certain variables were collected from a certain population.

Inclusion criteria:

- All race groups and both males and females were included in the study.

Exclusion criteria:

- Children less than 12 years old.

3.2.1 UM Wellness Health-Screening Programme

The University of Malaya (UM) is the leading university in Malaysia. It has more than 2,000 academic and 3,000 non-academic staff, with 17 faculties and 70 research centres that cover the whole spectrum of learning areas ("University of Malaya Official Portal,"). The UM Wellness Programme is an effort by the UM management in collaboration with the Department of Social & Preventive Medicine, Faculty of Medicine at this university, to promote the well-being of their employees' health. The UM Wellness Health-Screening Programme offers health screening or health-risk assessment which is able to identify employees who are at high risk of certain chronic diseases. Risk factors such as high cholesterol level, high blood pressure, obesity, physical inactivity, unhealthy diet, smoking and stress are identified, followed by health education and/or referral to a clinician or dietician where necessary.

The programme was established in response to the Malaysian government's policy introduced in the year 2003 (*Pekeliling 2003, Bil.3*)("Pekeliling Perkhidmatan Bilangan 3 ", 2003). This policy encourages all employees aged 40 years and above to conduct health screening and a physical examination at regular intervals. The UM Wellness Programme was launched successfully by the Deputy Minister of Higher Education, Dr Hou Kok Chung, on 24th June 2007 at the *Dewan Tunku Canselor*, University of Malaya. Since the launch, an annual health-screening programme has been conducted. The health screenings are conducted between the months of May and August every year. About 1300 to 1500 staff participate in the screening every year. Initially, the screening programme was offered to all staff aged 40 years and above, but since 2009, staffs aged 35 years and above have also been invited to the screening. Letters of invitation are sent out to all eligible staff for them to come within a specific time frame.

Data collection for this study was conducted between May and August 2009. The screening was conducted early in the morning before staff started their work. This was either between 7.30am and 8.30am, or between 8.30am and 9.30am. The screening programme was conducted every Tuesday to Friday morning at the Perdanasiswa Building, University of Malaya. Participants of this screening were both male and female staff from various racial backgrounds. Variables collected during from this population included blood glucose level, systolic blood pressure (SBP), diastolic blood pressure (DBP) and heart rate (HR).

3.2.2 UM Wellness Quit Smoking Clinic

The UM Wellness Quit Smoking Clinic was set up as part of the UM Wellness Programme starting from the year 2010. A total of 290 smokers attended this clinic. Services provided in the clinic include specific health assessment for smokers, individual and group counselling, and nicotine replacement therapy. This service is offered to all staff at all ages for free. All participants must come for the first initial session (for full health assessment and group counselling), and a fortnightly follow-up session depends on their time and progress.

For the year 2009, the UM Wellness Quit Smoking Clinic ran every Tuesday to Thursday between June 2009 and June 2010. The sessions for new cases were in the morning, and follow-up sessions in the afternoon. Each session took about 30-40 minutes. Invitation letters were sent out to all staff, but only for smokers to respond to (as the identity of smokers was unknown). Those who responded were invited to attend the clinic and participate in this study. Data collection for this population was obtained between June 2009 and April 2010. Data collected were body weight, body temperature, peak expiratory flow rate (PEFR) and carbon monoxide level (CO).

3.2.3 Community Health-Screening Programme, Klang, Selangor.

A community health-screening programme was organized by the researcher to assist with the data collection. This was also a social contribution by this research to the community. This screening programme was conducted in Kampung Teluk Gadong Kecil, Teluk Gadong, Klang, Selangor. Data collections for this population were performed between June 2009 and July 2009. This area was selected due to the presence of good local community contact in the area, and the wide range of age and socio-economic status of the population. Socio-economic status is well known to be associated with health status. This would give a good range of data for this study.

Kampung Teluk Gadong Kecil is located in the Teluk Gadong area in the Klang district of Selangor. Latitude and longitude are 3.0333° N and 101.4° E. The Teluk Gadong area was first opened in the mid 19th century. In 1910, Port Swettenham (now known as Port Klang) was opened by Sir Frank Swettenham ("Klang District Office," 2011; "Traditional Village in Selangor," 2010). Many locals and foreigners working in the port settled in new residence areas near the port, including Kampung Teluk Gadong Kecil, located about 5 km from Port Klang. This is one of many villages in the Teluk Gadong area.

The word “Teluk”, or “bay”, comes from the geographical location of the village on the outskirts of Sungai Klang (Klang river), and “Gadong” is derived from plants that grow in the bay area (Klang river) ("Traditional Village in Selangor," 2010). Thus, the name “Teluk Gadong” was coined by the early village settlers. The word “Kecil” means “small” in the Malay language. Estimated population for this village is about 1500 people and majority of the population are Malays ("Klang District Office," 2011; "Portal Rasmi Kampung Tradisional," 2012).

A total of eight health-screening sessions were set up by the researcher at a different location and time throughout the village. Information on the location and timing of each health-screening session was spread by a local representative to the community. Data collected from this population included heart rate (HR), systolic blood pressure (SBP) and diastolic blood pressure (DBP). A blood glucose level measurement using a glucometer was also offered during the session, for the benefit of the participants only and not included in this study.

3.2.4 Community Health-Screening Programme, Kuala Lumpur.

UM Community Health Awareness Day was organized by the University of Malaya and University of Malaya Medical Centre (UMMC) at the Mid Valley Exhibition Centre. The Mid Valley Exhibition Centre is located right on the top floor of the Mid Valley Megamall in Kuala Lumpur. This health screening was conducted on 6th and 7th February 2010 from 9am to 6pm. Data collection was performed at the Quit Smoking booth. Data collected were weight, body temperature, peak expiratory flow rate (PEFR) and carbon monoxide (CO) level. Both smokers and non-smokers of all ages (adults) and races were included in this study. This was to ensure a wide range of values for carbon monoxide levels.

3.3 Study Variables

Variables collected in this study were based on two sub-studies in this project (agreement study and reliability study).

3.3.1 Agreement Study

For agreement analysis, variables collected were:

1. Blood Glucose level: Laboratory value versus glucometer reading.
2. Systolic Blood Pressure (SBP): Manual sphygmomanometer SBP reading versus automatic SBP (first reading).
3. Diastolic Blood Pressure: Manual sphygmomanometer DBP reading versus automatic DBP (first reading).
4. Peak Expiratory Flow Rate (PEFR): Clement Clarke UK peak flow meter (first reading) versus Respicare peak flow meter.
5. Weight: Digital weighing scale versus analogue weighing scale.

3.3.2 Reliability Study

For reliability analysis, variables collected were:

1. Systolic Blood Pressure (SBP): First, second and third reading of automatic SBP machine.
2. Diastolic Blood Pressure (DBP): First, second and third reading of automatic DBP machine.
3. Heart Rate (HR): First, second and third reading of automatic blood pressure machine.
4. Body temperature (Temp): First, second and third reading of non-contact infrared thermometer.

5. Peak Expiratory Flow Rate (PEFR): First, second and third reading of Clement Clarke peak flow meter.
6. Carbon monoxide (CO) level: First, second and third reading of carbon monoxide meter (piCO Smokerlyzer).

3.4 Study Instruments and Procedure of Measurement

All instruments were brand new and specifically purchased for the purpose of this study except for the automatic blood pressure machine (OMRON HEM 907XL IntelliSense Professional Digital Blood Pressure Monitor) and the digital weighing scale (Seca 813 Robusta High Capacity Digital Floor Scale). Both of these instruments were borrowed from the UM Wellness programme and were calibrated before being used in the health-screening sessions.

3.4.1 Blood Glucose

The data for blood glucose level were used in the analysis of agreement only. Therefore, only two sets of measurement of blood glucose level were required (one from the glucometer and another from laboratory analysis). The results from the glucometer reading were compared with the results of blood glucose level obtained from the laboratory test.

1) Laboratory Blood Glucose Test

The data for laboratory blood glucose value were obtained from the UM Wellness Health-Screening Programme, where the blood samples were withdrawn simultaneously (just before) the measurement of blood glucose level using the glucometer. All blood taking for the blood glucose laboratory test were performed as part of the UM Wellness Health-Screening Programme. Consent for blood withdrawal (venepuncture) for the blood test was also gained as part of the screening programme. However, permission to access the blood results was obtained from the participant and the coordinator of the UM Wellness Programme.

2) Glucometer (Accu-Chek Advantage Meter model 2032)

The measurement of blood glucose level via glucometer was obtained using an Accu-Chek Advantage Meter model 2032 with Accu-Chek Advantage II test strip. The glucometer was used according to the manufacturer's guidelines ("Roche Accu-Chek Owner's Booklet," 2004). The code key was replaced and a control test was run every time a new box of test strips was used. A single-use lancet (Accu-Chek Safe-T-Pro Plus with Accu-Chek Softclix lancet device) was used to make a very small prick on the fingertip. The blood sample obtained from the fingertip was then dropped on the edge of the test strip (according to the user manual).

None of the blood glucose measurements from the glucometer readings performed in the community screening programme were included in this study. This was done only for the purpose of health screening (for the benefit of the participants).

3.4.2 Blood Pressure (Systolic and Diastolic)

A total of four sets of BP measurements were required for this study. For the agreement analysis, manual BP readings were compared with the first BP reading from automatic measurement. Three BP readings from automatic measurement were used for reliability analysis. All blood pressure measurements were performed from the same arm for each participant.

The first reading was taken using a manual sphygmomanometer, and the next three readings were taken using an automatic BP machine. All three BP readings were taken consecutively with a minimum of 15 seconds' interval between each reading. The study conducted by Yarrows et al. (Yarows, Patel, & Brook, 2001) has shown that a 15-second interval between blood pressure readings is as accurate as a one-minute interval.

1) Manual mercury sphygmomanometer (Desk type – Accoson)

The measurement of blood pressure using a manual sphygmomanometer was according to the Guide to Management of Hypertension 2008, developed by the National Heart Foundation of Australia (NHF, 2009). As recommended by the guideline, the results for systolic and diastolic BP were recorded to the nearest 2 mmHg. All manual blood pressure measurements, both in the UM Wellness Health-Screening Programme and community-based screening programme, were measured by the same researcher using the same sphygmomanometer.

2) Automatic Blood Pressure Machine (OMRON HEM 907XL IntelliSense Professional Digital Blood Pressure Monitor)

The automatic BP machine was used according to the user guide produced by the manufacturer ("Omron Instruction Manual," 2009). The machine was validated by the manufacturer and calibrated before it was used in the screening programme. The same

automatic BP machine was used in both the UM Wellness Health-Screening Programme and community-based screening programme (Kampung Teluk Gadong Kecil, Klang).

3.4.3 Heart Rate

The data for heart rate were only used for reliability analysis. The first, second and third readings were compared to assess the reliability. Therefore, only three sets of heart rate readings were required. All three sets of heart rate readings were taken from the same patient, and were measured using the same instrument.

Automatic Blood Pressure Machine (OMRON HEM 907XL IntelliSense Professional Digital Blood Pressure Monitor)

The heart rate measurement was obtained using the same automatic BP machine used for measuring BP. The automatic BP machine provided a heart rate reading for each time a BP measurement was taken.

3.4.4 Weight

Data for weight were used for analysis of agreement only. Therefore, only one set of weight readings was required from each weighing scale. The results from the analogue weighing scale were compared with the results of the digital weighing scale. The weighing scales were calibrated by the supplier before the start of the screening programme. The same observer took the reading measurements from the analogue scale throughout this study. Participants were weighed using the analogue scale first, before being weighed using the digital scale. Weight measurements for both participants from the UM Wellness Quit Smoking Clinic and community-based screening programme were taken using the same weighing scales. Both scales were placed on a flat, smooth and hard surface as instructed in the user manuals. During each measurement,

participants were asked to take off their shoes, remove heavy clothing (such as jackets) and empty their pockets. The measurement was taken with the participant standing still on the scale. The two scales used in this study were:

1) Digital weighing scale (Seca 813 Robusta High Capacity Digital Floor Scale)

This is an electronic flat scale with a maximum capacity of 200 kg ("Scales Galore: High capacity bathroom scales ", 2011). This scale shows the weight in kilograms (to two decimal places).

2) Analogue weighing scale (Hanson Weighing Machine H926)

This analogue scale has a maximum capacity of 130 kg ("Digital Scales Company: Hanson H926 mechanical bathroom scale," 2011). All readings were taken to the nearest kilogram by the same observer.

3.4.5 Peak Expiratory Flow Rate (PEFR)

A total of four sets of peak expiratory flow rate (PEFR) readings were required for this study. For the agreement analysis, the first PEFR readings from the Clement Clarke UK peak flow meter (as a reference or standard) were compared with readings from the Respicare peak flow meter. Three readings from the Clement Clarke UK peak flow meter were used for reliability analysis. Repeated readings were taken at 1-2 minute intervals. All PEFR readings for both peak flow meters were taken by the same researcher. Techniques for measuring PEFR for both peak flow meters were as suggested by the manual from the Asthma Center Education and Research Fund (Dunsky et al., 2005):

1. Connect a clean mouthpiece.
2. Ensure the marker is set to zero.

3. Participant has to sit upright.
4. Ensure that the participant's mouth is empty.
5. Take a deep breath in and hold the breath.
6. Place the mouthpiece in the mouth.
7. Keep fingers away from the marker and vents of the meter.
8. Form a seal as tight as possible around the mouthpiece with lips.
9. Breathe out as hard as possible.

1) Mini Wright Standard Peak Flow Meter Clement Clarke UK

This is the original portable peak flow meter and the standard used by the majority of health-care professionals ("Mini-Wright Standard," 2011). The Mini Wright Standard Peak Flow Meter has a scale ranging from 60 to 800 litres per minute. This meter was used with disposable cardboard mouthpieces, which were replaced for each new participant.

2) PulmoPeak Peak Flow Meter, Respicare

The PulmoPeak peak flow meter comes with a scale ranging from 60 to 900 litres per minute. It incorporates the popular zoning system, which is particularly useful in asthma management (green, yellow and red zones). According to the American Lung Association, the three-zone system will help doctors and health practitioners develop an asthma management plan for their patients ("Take Control of Your Asthma," 2012). A peak flow reading in the green zone indicates that the asthma is under good control. A peak flow reading in the yellow zone indicates caution is necessary, and this may mean that additional medication is required. Finally, a peak flow reading in the red zone indicates a medical emergency. This usually suggests that immediate action needs to be taken.

3.4.6 Body Temperature

Three sets of temperature readings were required for the reliability analysis in this study. All temperature readings for both the UM Wellness Quit Smoking Clinic and community-based screening programme were measured using the non-contact forehead infrared thermometer by the same researcher.

Non-contact Forehead Infrared Thermometer (DT-8806H)

All measurements were taken from the patient's forehead, with a measurement distance between 5 cm and 15 cm (according to the specifications of the manufacturer) ("IR Thermometer (DT-8806H)," 2009). Second and third readings were taken approximately from the same distance and after about a three-second interval. The response time of this thermometer is about 0.5 seconds ("IR Thermometer (DT-8806H)," 2009). Since this thermometer has two temperature settings (body and surface temperature), body temperature setting was set throughout all the measurements.

3.4.7 Carbon Monoxide (CO) Level

Data for carbon monoxide level were only used for analysis of reliability. Therefore, three repeated measurements of CO level were performed for the same participant using the same instrument.

Smokerlyzer (piCO Smokerlyzer)

The carbon monoxide level (CO) was measured using the piCO Smokerlyzer. The reading was given in COppm, which was the number of CO molecules in a million parts of air. According to the specification of the manufacturer, this meter is able to detect the concentration of CO level between 0 and 80 ppm, with a response time of less than 45 seconds (*piCO+ Smokerlyzer User Manual*, 2006).

Measurement technique was according to the user manual (*piCO+ Smokerlyzer User Manual*, 2006). The start button on the meter was clicked twice to start the breath test. The participant was asked to inhale and hold their breath for 15 seconds. The double clicks of the button at the start of the test initiate a 15-second countdown. The meter produced a warning bleep sound during the last three seconds of the countdown. Then the participant was asked to blow slowly into the mouthpiece aiming to empty their lungs completely.

Participants blew into the meter through a single-use disposable cardboard mouthpiece, which was connected to the meter via a unique breath-sampling "D-piece". The D-piece has a special feature (a one-way valve) to prevent air from being drawn back through the monitor (*piCO+ Smokerlyzer User Manual*, 2006). The D-piece also has an infection control filter, which filters out most airborne bacteria. The D-piece was shown to remove and trap about 99.9% of airborne bacteria (*piCO+ Smokerlyzer User Manual*, 2006). This system protects the instrument from contamination, and also reduces the risk of cross infection among participants.

3.5 Ethical Approval and Funding

As part of the requirement for clinical research involving humans, the principal investigator for this study had to attend a Good Clinical Practice (GCP) course and pass the examination in this course. After receiving the GCP certificate, applications for ethical clearance and funding were submitted.

There was no problem with the process of ethical clearance application. In April 2009, this study was approved by the Medical Ethics Committee of the University of Malaya Medical Centre (MEC ref no: 715.23), without any revision of the proposal and any defence to the ethical committee. All the major work for the data collection started as soon as the approval was received.

Finally, in August 2009, approval for funding was received. This study was fully funded by a Postgraduate Research Grant (PPP) provided by the Institute of Research Management and Monitoring (IPPP), University of Malaya (grant number: PS162/2009B), and a High Impact Research (HIR) Grant (UM/MOHE) (grant number: E000010-20001). The total amount of grant received for this project was RM20,024. This amount was allocated to a number of expense categories.

Table 3.1: Budget allocation

GRANT	EXPENSE CATEGORY	ALLOCATION (RM)
IPPP grant	Conference/Symposium	3,253.20
	Consumables	4,377.00
	Equipment	4,410.00
	Salary/honorarium	3,450.00
	Travel	419.80
HIR grant	Publication	4,114.00
	TOTAL	**20,024.00**

3.6 Sample Size Calculation

Calculation of sample size was performed based on the statistical analysis required for this study. The main statistical analysis that is going to be run under this study includes linear regression analysis (comparison of slopes and y-intercepts analysis), the Bland-Altman method (Limits of Agreement) and the Intra-class Correlation Coefficient. The sample size estimation for the Limits of Agreement (LoA) and the ICC depends on the precision of prediction. Higher precision prediction will require a bigger sample size. However, by estimating the confidence intervals of the Limits of Agreement, Bland (Bland, 2004) recommended a sample size of 100 for the Bland-Altman analysis.

A formal sample size calculation was performed based on Cohen's statistical power analysis (J. Cohen, 1988). The calculation was according to the formula based on the method suggested by Cohen in 1977. Cohen suggested power tables for the determination of power for the multiple-regression analysis based on the function of L (the non-centrality parameter) (J. Cohen, 1977).

$L = f^2 v$

v = the error (denominator) = N – u – 1

u = the numerator = k – 1

f^2 = effect size

N = total sample

k = number of group

The estimation of sample size was based on the alpha level of 0.05, with desired statistical power of 80%, and anticipated effect size (f^2) of 0.02. The effect size of 0.02 was considered to be a small effect size (J. Cohen, 1977). In this study, there will be two methods of measurement per variable (k = 2), where each method is applied to n

patients (so total N = 2n). From the power table (Appendix K) proposed by Cohen (J. Cohen, 1977), when u = 1 and the power of the study is 80%, the value of L = 7.9.

From the formula $L = f^2v$:

$$v = \frac{L}{f^2} = \frac{7.9}{0.02} = 395$$

$$v = N - u - 1$$

$$395 = N - 1 - 1$$

$$N = 397$$

$$n = 198.5 \approx 199$$

The number of the sample size should be at least 199. However, for the purpose of analysis, 300 measurements were collected for each variable. Retrospectively, this would give a power of 93%.

3.7 Data Collection

Written informed consent was obtained from all the participants. All participants had the choice whether to join the study or not. The purpose of this study and the need for repeated measurements in this study were explained to the participants. Although repeated measurements were not beneficial to the participants and may cause some discomfort, it is unlikely that it will jeopardize their health at any time of their life. This issue was explained to all participants as part of the informed-consent process. Participants were also permitted to withdraw from this study at any time during the study period. However, none of the participants asked to withdraw.

The main work for data collection started on 22nd April 2009 after the ethical approval was obtained. This began with the designing and printing work (for the patient information sheet, health information sheet and consent form), contacting community representatives, and getting supplies for all required equipment and necessary stationery. Data collection was divided into two phases.

3.7.1 Phase I

Phase I was conducted between May and August 2009. Variables collected during this phase included Blood Glucose level, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP) and Heart Rate (HR).

3.7.1.1 UM Wellness Health-Screening Programme

Data were first collected on 12th May 2009 during the UM Wellness Screening Programme. All 300 blood glucose samples were obtained during this screening programme, because blood glucose values from laboratory analysis were only available during this screening programme and not in the community screening programme.

The initial plan was to collect all glucose level readings from the UM Wellness Screening Programme. For blood pressure and heart rate readings, the plan was to collect 150 samples (for each data set) from the UM Wellness Screening Programme, and another 150 from the community. For UM Wellness data collection, all measurements of SBP, DBP, HR and glucose were planned to be taken at the same time as seeing the participant. However, during the first week of UM Wellness screening, glucose strips were not available as there was a delay in the delivery by the supplier. So during the first week, only data for blood pressure and heart rate were collected. After experiencing the first week of data collection, during which BP measurements (four readings) for each participant required at least 10 minutes (including consent procedure), on average only 10 participants were obtained per session. Hence, there was a change in plan.

Data for glucose were given priority as this had to be completed throughout the UM Wellness screening due to the necessity for the laboratory results for comparison. As soon as the glucose strips were available, on 20th May 2009, only data for glucose

were collected until all 300 samples were collected. Initial concern about the willingness of the participants to give consent for another blood glucose reading (as they had already given blood for the laboratory glucose test) was not an issue during data collection. Most of the participants were in fact keen to know their blood glucose result on the spot. Only a very few participants refused to participate.

Although only 300 were required for this study, a total of 315 blood samples for the glucometer test were collected to make up for any missing data from the glucose laboratory result. Data collection for glucose from the glucometer test was completed on 5th June 2009. A total of 300 samples with matching glucose lab results were finally obtained. After that, data collection for BP and HR was then continued throughout the UM Wellness Screening Programme, which ran from 12th June to the end of August 2009.

3.7.1.2 Community Data Collection (Klang, Selangor)

Concurrent with the data collection at the UM Wellness Screening Programme, data collection for blood pressure and heart rate was also started in the community on 10th June 2009. The study area was the village of Telok Gadong Kecil in the Klang District, Selangor. Initially, data collections were performed from house to house. However, this was quite challenging, and time-consuming. After discussion with the local representatives, they agreed to set up a small centre and invite a small group of people in one session. The reception from the community was overwhelming, and eventually a total of eight small centres at different locations were set up for different sessions throughout the data collection in the community.

Although only BP data were required, blood glucose measurement was performed, and basic medical advice was given as a service to the community. Light refreshment was also provided to the participants as a token of appreciation and because

some of them were fasting for the blood glucose test. Consent for children below 18 years old who participated in this study was directly obtained from the children (not the parents). Gillick competency (Hunter & Pierscionek, 2007) was applied as the research is likely to be beneficial, while exposing the children to relatively very small risks. Only blood pressure measurement, and no blood glucose measurement, was performed for children below 18 years old.

It was an exceptionally satisfying experience to be in the community. The community gave a positive reception to the health-screening session and appreciated the opportunities. Some of the participants were retired elderly people and had never had any regular health screening. The pace of the health screening was quite slow (at least 15 minutes per participant) due to a lack of assistants and repeated BP measurements, and some of the participants took the opportunity to consult about their medical problems. However, none of the participants issued any complaint or appeared to suffer discomfort having to wait for the health-screening services. In fact, most of them welcomed the very convenient health services near to their door step and expressed gratitude. The last data collection for BP in the community was on 24th July 2009 with an overall total of 193 records for BP and HR collected.

3.7.2 Phase II

Phase II of the data collection was conducted between October 2009 and March 2010. Variables collected during this phase included body temperature (Temp), weight, peak expiratory flow rate (PEFR) and carbon monoxide level (CO).

3.7.2.1 UM Wellness Quit Smoking Clinic

The preparation for the data collection started with the setting up of the venue for the clinic itself. The year 2009 was the first year that the University of Malaya had run the Quit Smoking Clinic (with free consultation and treatment) for their staff. As there was no dedicated place for this, the clinic and its facilities had to be located in a vacant room. The room was located at Level 3, Block F, Perdanasiswa Building, University of Malaya. The setting up of the clinic took about two weeks and was finally completed on 26th October 2009. Data collection started on 27th October 2009.

For the first few weeks the participations in the clinic were terribly slow, with an average of three new patients per day. However, on entering the second week, the participation for the smoking-cessation clinic improved. The impact of the Quit Smoking programme was also remarkably good. Almost all the participants in the follow-up session had reduced the number of cigarettes smoked, and some participants had stopped smoking completely. The "snowball" effect among the smokers increased the number of participants for the clinic. The progress of data collection then continued to improve steadily, and by the end of November 2009, the total sample collected had reached 130.

However, after January 2010 the progress of data collection became slow again because most of the smokers who were willing to participate had already attended the clinic. Some of the patients were also discharged from the clinic because they had successfully quit smoking, and there were also patients who failed to attend the follow-up session. There were very few new cases every session (an average of two per day). Finally, a total of 204 readings for each variable were collected within this population.

3.7.2.2 Community Data Collection (UM Community Health Awareness Day)

Another data collection exercise from the community was conducted on 6th and 7th February 2010 from 9am to 6pm at the Mid Valley Exhibition Centre (Mid Valley Megamall, Kuala Lumpur). The data collection was performed as part of the activities during the UM Community Health Awareness Day at the Quit Smoking booth. The progress of data collection on both days was very good, due to assistance from some of the volunteers, and less consultation was performed in comparison to the sessions in the UM Quit Smoking Clinic. Participants were from various backgrounds and races and included both genders. A total of 96 readings for each variable was collected during these two days.

3.7.3 Summary of Data Collection

Proper planning, organization and time management, among other factors, contribute to the completion of data collection. However, the duration of data collection was longer than expected. Overall, the duration for the data collection was about four months for Phase I and six months for Phase II. This was due to a number of constraints during the process of data collection. The main constraints were a lack of assistants, a rigid schedule for working together with the UM Wellness Screening Programme, and slow participation for the UM Wellness Quit Smoking Clinic. Nonetheless, the opportunities available during the UM Wellness Screening Programme made the process of data collection a success. The facilities and some assistants during the programme helped the process of data collection run reasonably well.

3.8 Software Tools

The following software tools were used for the project:

1) SPSS Version 17.0

SPSS (Statistical Package for the Social Sciences) is a statistical package for data analysis. This statistical software can be used to perform highly complex data manipulation and analysis with simple instructions. Many features of SPSS are accessible via a drop-down menu or can be programmed with a syntax command language. All raw data were entered into this software. Data cleaning and all descriptive analyses were performed using this software. Correlation, linear regression and Intra-class Correlation Coefficient analyses were also performed using this software.

2) GraphPad Prism 5.02

GraphPad Prism combines scientific graphics, statistics and curve fitting functions. This software was originally made for biological studies, but now includes several sophisticated non-linear capabilities as well as scientific graphing that would be useful to other scientists. This software has simple instructions, which makes it user-friendly software. Technical explanation of analysis in this software is also provided. This software was used to perform the Bland-Altman analysis and comparison of slopes and y-intercepts analysis because it has a special feature to perform this analysis.

3) Matlab Programme Version 7.8 (R2009a)

Matlab is a high-performance language for technical computing. The name Matlab stands for “matrix laboratory”. It integrates computation, visualization and programming. Most functions in this software require syntax command language. This

software was used mainly for data simulation and analysis to see the effect of sample size for all tested methods.

4) MedCalc Statistical Software Version 12.1.3

MedCalc statistical software is designed to help biomedical researchers in performing statistical analyses. It comes with a large number of tools and tests, as well as a number of statistics and calculi that it can carry through. This software is user-friendly and not difficult to operate. In this study, this software was used to test for proportional bias in the Bland-Altman analysis.

5) Microsoft Excel 2007

Microsoft Excel forms part of Microsoft Office. It has the basic features of all spreadsheets using a grid of cells to organize data manipulations. This software also has graphical and basic mathematical and statistical analysis functions. It was used for data transfer.

3.9 Statistical Methods Used

This section describes some of the statistical concepts and different statistical methods that have been used in this study. The statistical methods used for assessment of agreement and reliability in this study are also explained later in this section. There are many statistical methods that have been used to assess agreement and reliability. However, the statistical methods chosen for analysis in this study were the most commonly used methods in the medical literature, as shown in the systematic review carried out earlier (Chapter 2).

3.9.1 General Statistical Concepts and Methods Used

3.9.1.1 Normal Distribution

One of the most important theoretical distributions in statistics is the *normal distribution* (Daly & Bourke, 2000). The normal distribution is important because it is a good empirical description of many variables, and it occupies a central role in the techniques of statistical analysis (Kirkwood, 2000).

This distribution is also known as *Gaussian distribution*, named after Carl Friedrich Gauss (1777-1855) a German mathematician, after he discussed the *normal distribution* law in 1809 (Zar, 2010). The distribution was actually first used by de Moivre in 1738 to approximate a binomial distribution (Zar, 2010). However, it was Pierre-Simon Laplace, a French mathematician and astronomer, who proved and emphasized the theoretical importance of the *normal distribution* in 1810 (Pearson, 1905). The adjective *normal* was first used for the distribution by Charles S. Peirce in 1873, and the use of the term "normal" was recommended to avoid "an international question of priority" (Zar, 2010).

The normal distribution is defined by its probability-density function (Rosner, 2006). The density function follows a bell-shaped curve. The curve is symmetrical about the mean (μ) with the mode at mean. The shape of the normal distribution is determined by the mean (μ) and variance (σ^2). The curve is tall and narrow for small variances, and short and wide for large variances. The normal distribution is usually denoted by N (μ,σ^2). A normal distribution with mean 0 and variance 1 is called a *standard normal distribution*, and referred to as an N (0,1) distribution (Rosner, 2006).

The term "normal" does not mean that the distribution is common or typical. In fact, most variables in medical research are non-normal, and this is actually not an abnormal situation (Daly & Bourke, 2000). The occurrence of normal distribution in a practical situation can be loosely classified into three categories: exactly normal, approximately normal, and distribution modelled as normal. An exactly normal distribution for a variable is not actually a prerequisite for many forms of statistical analysis, but an *approximately* normal distribution was in fact required in many situations (Daly & Bourke, 2000). Many random variables in the general population, such as blood pressure and weight, tend to follow approximately a normal distribution. Any random variables with non-normal distribution can be made approximately normal by transforming the data onto a different scale.

Most estimation procedures and hypothesis tests assume that the random variable being considered has an underlying normal distribution. Tests will not be valid if this assumption is violated. Therefore, it is important to examine shapes of distribution before applying any statistical analysis. There are numerous tests for normality, such as the Q-Q plot, the Shapiro–Wilk test, the D'Agostino's K-squared test and the Kolmogorov–Smirnov test. In general, normality tests assess the likelihood that the given data set comes from a normal distribution (Razali & Wah, 2011).

The normality of a distribution always refers to the underlying distribution of a variable in a population. If random samples of size *n* are drawn from a normal population, the distribution of means for these samples will be normal. The distribution of means from a non-normal population will not be normal, but will tend to approximate to a normal distribution as *n* increases in size. This result is known as the *Central Limit Theorem* (Zar, 2010).

3.9.1.2 Central Limit Theorem

The *Central Limit Theorem* states that the sampling distribution of any statistic will be normal or nearly normal, if the sample size is large enough (Kirkwood, 2000). The number needed to give a close approximation to normality depends on how non-normal the population is (Kirkwood, 2000).

The advantage of the *Central Limit Theorem* is that sample data drawn from populations of unknown shape or not normally distributed can also be analysed using a statistical test with normal distribution assumption. This is because the sample means are normally distributed for sample sizes of $n \geq 30$ (Arjomand, 2002). Figure 3.1 shows the shape of the distribution of the sample means for a particular sample size of four different population distributions. The distribution of the sample means begins to approximate the normal curve as the sample size *n* increases (Arjomand, 2002).

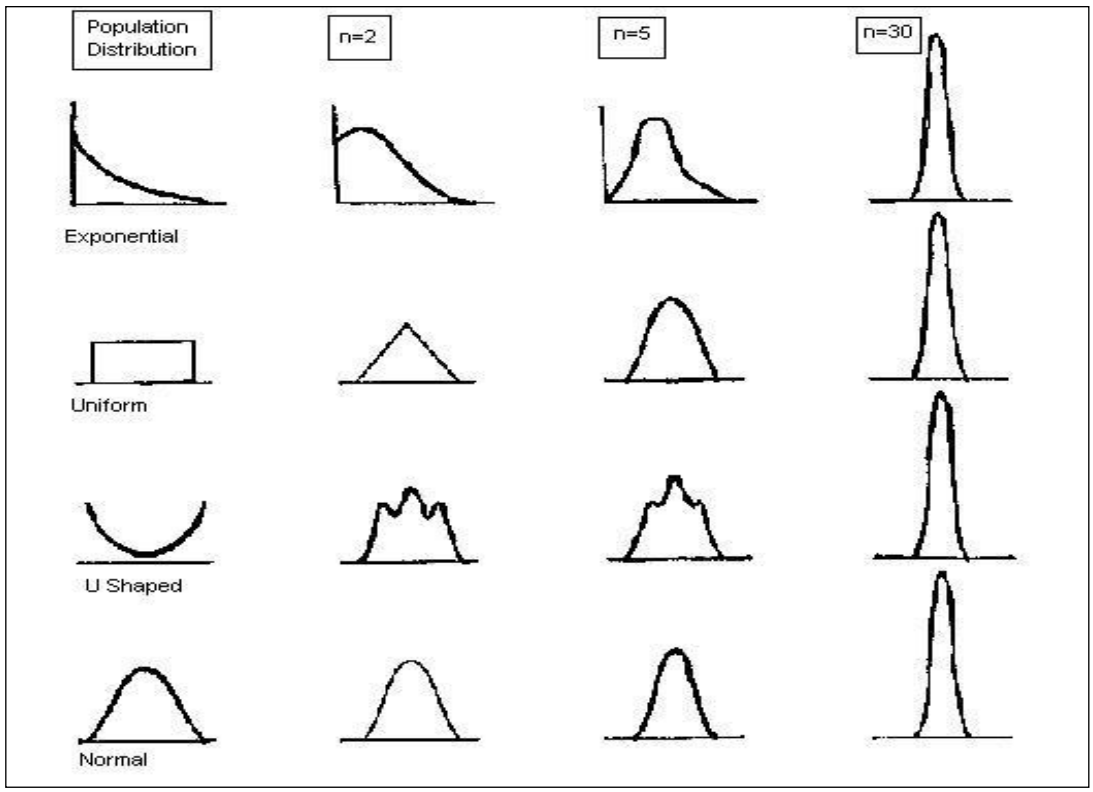


Figure 3.1: The distribution of the sample means for different population distributions (Arjomand, 2002)

3.9.1.3 Simple Linear Regression

Linear regression describes the relationship between two continuous variables. Linear regression gives the equation of the straight line that describes how the "*y"* variable changes (increases or decreases) with changes in the "*x"* variable. The equation of the regression line is given as $y=\alpha+\beta x$. "*y"* is the dependent variable and "*x"* is the independent or explanatory variable, where "α" is the y-intercept and "β" is the slope of the line. The slope "β" is also sometimes called the *regression coefficient* (Kirkwood, 2000)

There are a number of ways of fitting the linear regression, such as the ordinary least-squares (OLS) method and ordinary least-products (OLP) method. There were debates about which methods should be used to fit the regression line in relation to the analysis of agreement. The application of both OLS and OLP methods in the analysis of

agreement was discussed earlier in Chapter 2. In this study the analysis of agreement based on the comparison of slopes and y-intercepts analysis will be based on the OLS method. For the OLS method, the values for "α" and "β" are calculated to minimize the sum of squares of the vertical deviations of the points about the line (Kirkwood, 2000).

There are certain principal assumptions that must be met which justify the use of linear regression models (Zar, 2010):

(i) **Linearity** – The actual relationship between dependent and independent variables is linear.

(ii) **Homoscedasticity** – The variances of distribution of *y* values must all be equal to each other (homogeneity of variances in the population).

(iii) **Independence** – For each value of *x*, the values of *y* are to come at random from the sampled population, and are to be independent of one another.

(iv) **Normality** – For any value of *x* in the population, there exists a normal distribution of *y* values.

(v) The measurements of *x* were obtained without error – this is almost impossible, so it is assumed that the errors are very small.

Testing assumption

It is important to assess these assumptions, because if any of these assumptions is violated, then the prediction yielded by a regression model may be biased or misleading. Since the sample size in this study is large, the data were assumed to be normally distributed (this is according to the Central Limit Theorem). However, the skewness and kurtosis of the data were assessed in the descriptive analysis. Assumption number (iii) on the independence of the response variables is subject to the design of the study, and

the way the data have been collected (Larsen, 2008). In this study (for agreement analysis), all data have been collected independently; therefore this assumption is satisfied. Assumption number (v) was also not a big problem in this study. To satisfy this assumption, measurements of *x* were obtained from a standard instrument. Therefore only assumptions of linear relationship and homoscedasticity need to be checked. Residual analysis can be used to check these assumptions:

(i) **Linear Relationship –** This assumption can be assessed using a residual plot. A systematic pattern of the residuals suggests that this assumption is violated (Chambers, 2008).

(ii) **Homoscedasticity –** In a residual plot, the residuals should appear random with constant variance. If there appears to be a change in the variance, then the assumption of constant variance in the residuals may be violated (Chambers, 2008).

3.9.2 Method Used to Assess Agreement

3.9.2.1 Comparison of slopes and y-intercepts

A simple linear regression method is the proposed method of measuring agreement. This consists of a comparison of slopes and y-intercepts and agreement model. GraphPad Prism software and SPSS software were used to perform the analysis. The first step for the analysis was to obtain the two linear regression equations:

a. Two linear regression equations:

1. **Line of $y = \alpha + \beta x$ (tested line)**

   $\boldsymbol{x}$ = value from standard instrument (e.g. blood glucose level obtained from laboratory)

   $\boldsymbol{y}$ = estimated value using alternative instrument (e.g. blood glucose value from glucometer)

   $\boldsymbol{\alpha}$ = intercept

   $\boldsymbol{\beta}$ = slope

2. **Line of agreement, $y_2 = x$**

   It was assumed that the estimated value "y_2" is exactly the same as the value from standard instrument "x". Therefore, during the analysis this line was obtained by plotting the value of y_2 (which is exactly the same value as x) against the value x.

If there is an agreement between y and x, the line of $y = \alpha + \beta x$ will be as close as possible to the line of $y_2 = x$. In other words, there will be no difference between the two lines. To achieve this, the slope β has to be very close to 1 or equal to 1, and the

intercept α has to be very close to zero or equal to zero. However, in testing two instruments, what is important is the range of differences between the two instruments.

b. Testing linearity

The relationship between *y* and *x* has to be linear for *y* and *x* to be in a good agreement. It is impossible for *y* and *x* to have an agreement if their relationship is non-linear. The linearity of the line of $y = \alpha + \beta x$ was assessed using Pearson Correlation Coefficient (*r*) before comparing it with the line of agreement. The correlation coefficient $r \geq 0.8$ was considered to be a strong linear relationship (Chan, 2003).

c. Comparing slopes

The next step was to compare slopes of line $y = \alpha + \beta x$ and line of agreement $y_2 = x$. Detail of the formulation was explained in Chapter 2.

The GraphPad Prism software was used to calculate a *p*-value (two-tailed) testing the null hypothesis that both the slopes are identical (the lines are parallel). This software used F-test to test the null hypothesis (method described in Section 2.5.1.1) (Motulsky, 2007). If the *p*-value is less than 0.05, the lines are significantly different. In that case, there is no point in comparing the intercepts. However, if the *p*-value is greater than 0.05, the slopes are not significantly different, and the next step was to compare the intercepts.

d. Comparing intercepts

If the slopes are indistinguishable, the lines could be parallel with distinct intercepts, or the lines could be identical with the same slopes and intercepts. The GraphPad Prism software was used to calculate a second *p*-value testing the null hypothesis that the intercepts of both lines are identical (using F-test) (Motulsky, 2007). Detail of the

formulation was explained in Chapter 2 (Section 2.5.1.2). If this *p*-value is greater than 0.05, there is no compelling evidence that the lines are different. However, if this *p*-value is less than 0.05, the lines are not identical (they are distinct but parallel).

e. Residual analysis

The residual is the difference between an observed value and predicted value from the regression line.

Residual = Observed value – Predicted value

A small residual means that the prediction value is close to the observed values. The main purpose of performing this analysis is to assess the linear regression assumption. Using SPSS software, the residuals of the regression line were plotted against the standard values, to assess the pattern of residuals. The normal distribution was approximated by plotting a histogram, and calculating the skewness and kurtosis. Figure 3.2 shows an example of a residual plot.

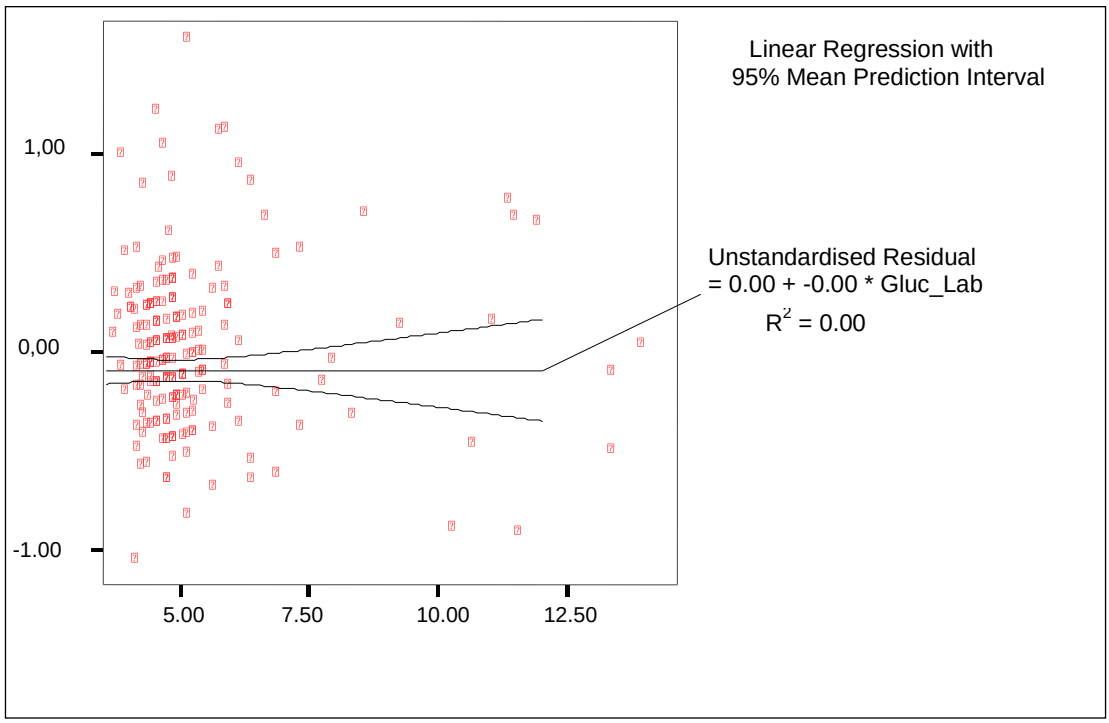


Figure 3.2: Example of residual plot

3.9.2.2 Agreement Model

In a method comparison study, a certain range of differences of measurements will be accepted as an agreement. Using the linear regression equation $y = \alpha + \beta x$ and the function of error equation $f\ (Error) = Predicted\ value\ (y) - True\ value(x)$, an agreement model is proposed to quantify bias produced by an instrument.

$$Error = y - x$$

$$y = \alpha + \beta x$$

$$Error = \alpha + \beta x - x$$

$$Error = \alpha + (\beta - 1)x$$

Mean, minimum and maximum errors were estimated using the mean, minimum and maximum value of *x* (value obtained using standard instrument) in the data set. The estimated error was compared with the significant or tolerable clinical difference value (see Table 3.2). If the estimated error was less or equal to the tolerable clinical difference value, this means that there was an agreement between the two instruments. The tolerable clinical difference values for glucose level (Essack et al., 2009), systolic and diastolic BP (Pickering et al., 2005), peak expiratory flow rate (Quanjer PH, Lebowitz MD, Gregg I, Miller MR, & Pedersen OF, 1997) and body weight ("National Weights and Measures Laboratory," 2003) used in this study were obtained from previous studies. The values are displayed in Table 3.2.

Table 3.2**:** Tolerable clinical difference value set in this study

Variable	Standard Instrument	Alternative Instrument	Clinical difference
Glucose	Laboratory value	Glucometer	±0.8mmol/l (Essack et al., 2009)
Systolic Blood Pressure, SBP	Manual sphygmomanometer	Automatic BP machine	±10mmHg (Pickering et al., 2005)
Diastolic Blood Pressure, DBP	Manual sphygmomanometer	Automatic BP machine	±10mmHg (Pickering et al., 2005)
Peak Expiratory Flow Rate, PEFR	Clement Clarke UK peak flow meter (first reading)	Respicare peak flow meter	±40 l/min (Quanjer PH et al., 1997)
Body Weight, Wt	Digital weighing scale	Analogue weighing scale	±0.5kg ("National Weights and Measures Laboratory," 2003)

To demonstrate the method, Table 3.3 shows hypothetical data for blood glucose measurement using glucometer A and compared with the laboratory values. The linear regression equation of y against x is $y = 1.988x - 0.988$, so from the agreement model,

$$Error = \alpha + (\beta - 1)x$$

$$Error = (1.988 - 1)x - 0.988$$

$$Error = 0.988x - 0.988$$

Mean value of glucose level for these data is 8.7mmol/l, so most of the time the glucometer will produce a positive error of 7.6mmol/l. The slope of the line suggests that the error increases with the increased value of glucose level. The minimum value of glucose in the data set is 7mmol/l and the maximum value is 11mmol/l. So, the

minimum error will be 5.9mmol/l and the maximum error will be 9.9mmol/l. Since the maximum error is more than the tolerable clinical difference value (0.8mmol/l), there is no agreement between the glucometer reading and glucose level from the laboratory analysis in Table 3.3.

Table 3.3: Hypothetical data of blood glucose measurements

Lab (mmol/l)	**Glucometer A (mmol/l)**	**Predicted Error (mmol/l)**
10.0	20.0	8.9
8.0	16.0	6.9
8.0	16.0	6.9
10.0	19.0	8.9
9.0	18.0	7.9
8.0	16.0	6.9
9.0	18.0	7.9
7.0	13.0	5.9
7.0	13.0	5.9
11.0	21.0	9.9

Summary of proposed method:

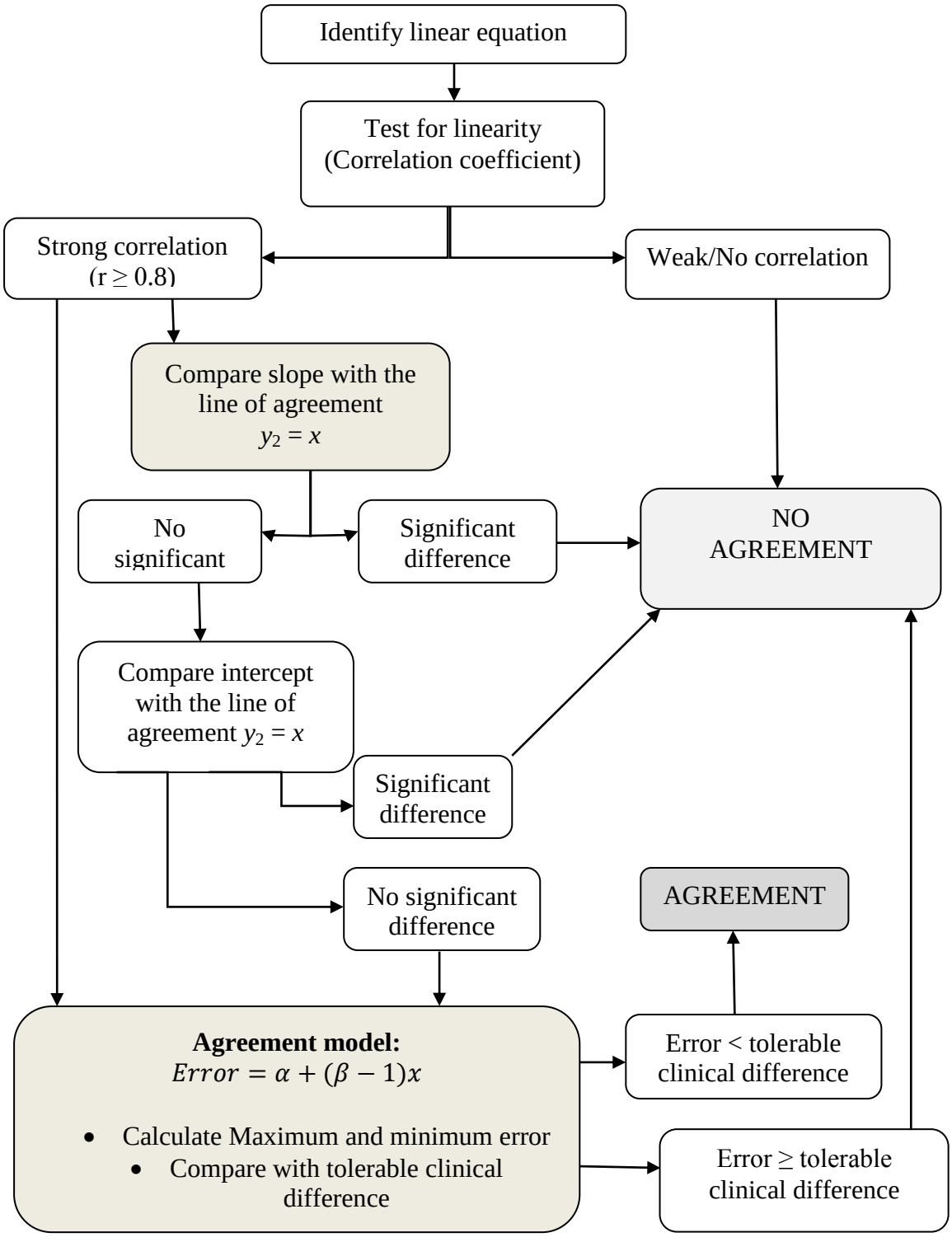


Figure 3.3: Proposed method of assessing agreement

3.9.2.2 Bland-Altman Method (Limits of Agreement, LoA)

The Bland-Altman analysis (Limits of Agreement, LoA) was performed using the GraphPad Prism software. The formula for Limits of Agreement is given as (Bland & Altman, 1987):

$$LoA = mean\ difference\ \pm 1.96 \times standard\ deviation\ of\ the\ differences$$

To ensure the assumption for the Bland-Altman analysis was not violated, the pattern of bias (differences) was observed using the Bland-Altman plot and the distribution of the differences was evaluated by plotting the histograms of the differences of two measurements, and the D'Agostino and Pearson omnibus normality test. The D'Agostino-Pearson test first analyses the skewness and kurtosis of data, and then calculates how far these values differ from the expected values for a normal distribution (Oztuna, Elhan, & Tuccar, 2006). This test tends to be more powerful than the Kolmogorov-Smirnov test (Oztuna et al., 2006).

The limits of agreement were compared with the tolerable clinical difference value (see Table 3.2). If the limits of agreement were less or equal to the tolerable clinical difference value, this means that there was an agreement between the two instruments.

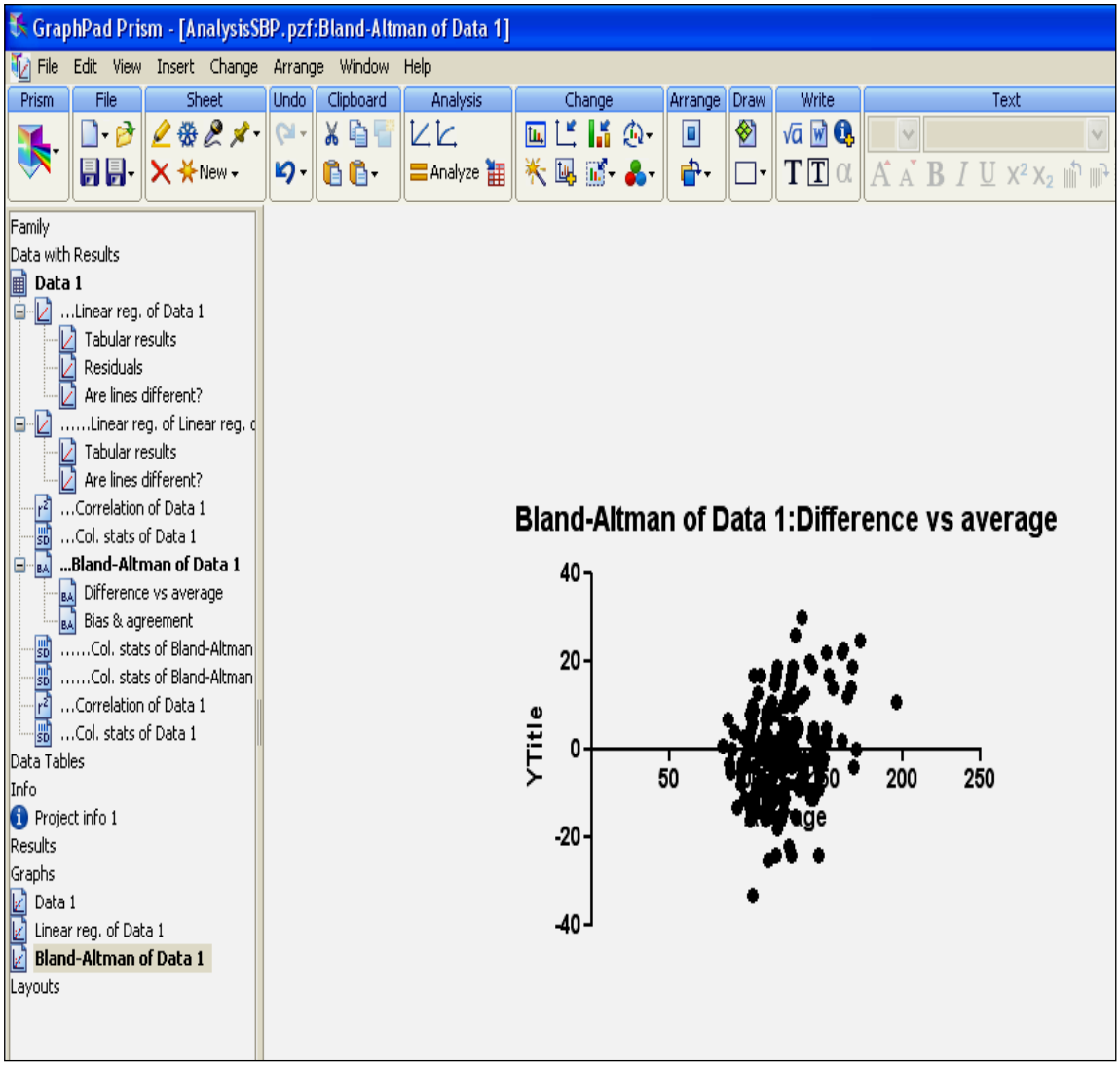


Figure 3.4: Example of Bland-Altman plot

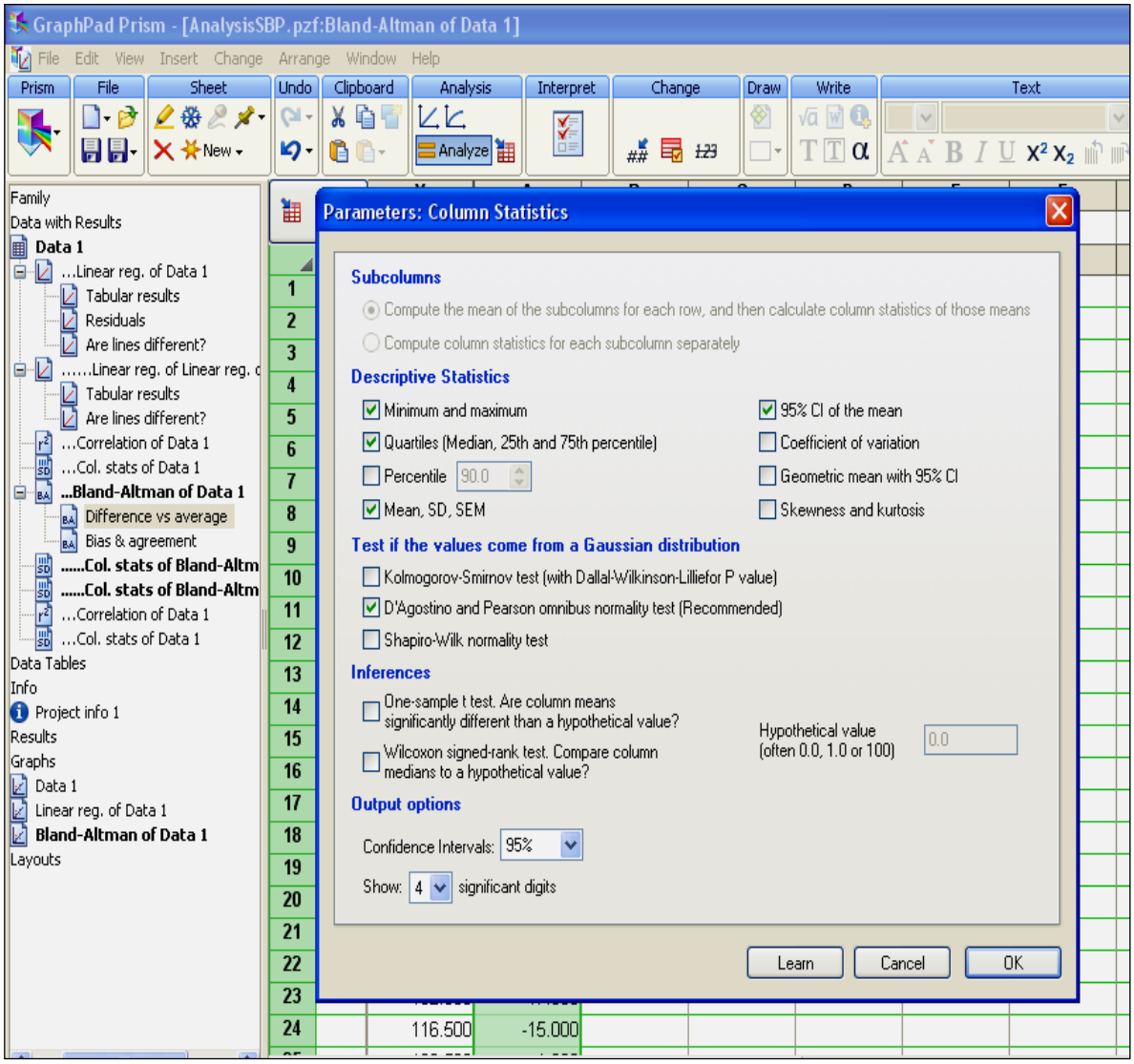


Figure 3.5: Testing normality using GraphPad Prism software

3.9.2.3 Intra-class Correlation Coefficient

The Intra-class Correlation Coefficient for agreement (ICC_A) analysis was performed using SPSS software, under reliability analysis with absolute agreement option. Since the purpose of the analysis was to test agreement, a two-way random model was chosen (Weir, 2005). Different models for the ICC and mathematical formulations were discussed in Chapter 2. In this model, it is assumed that each subject was assessed by the same raters, where these raters were randomly sampled from the population (raters are considered as a random effect) (Weir, 2005). In this study, the value of $ICC_A \geq 0.75$ was considered to be good agreement (Rosner, 2006).

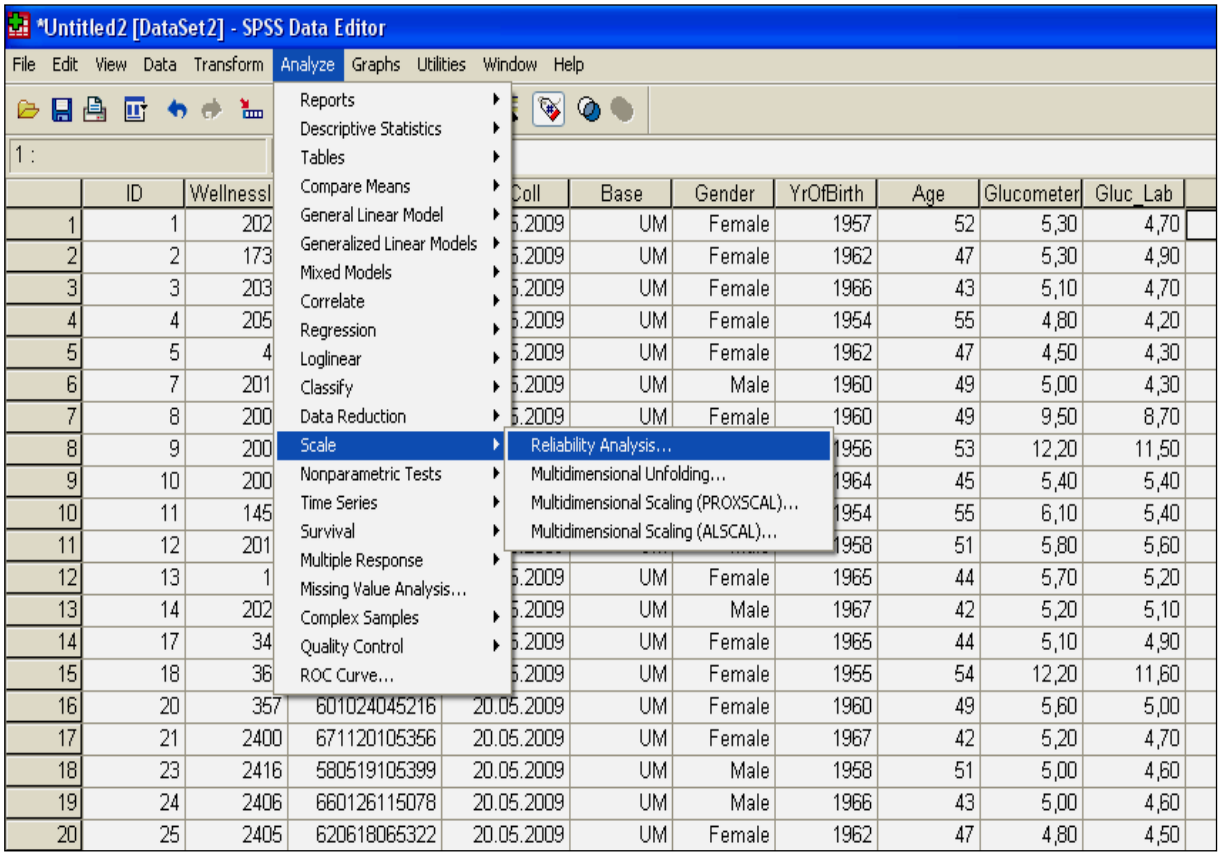


Figure 3.6: Analysis of ICC_A for agreement using SPSS (a)

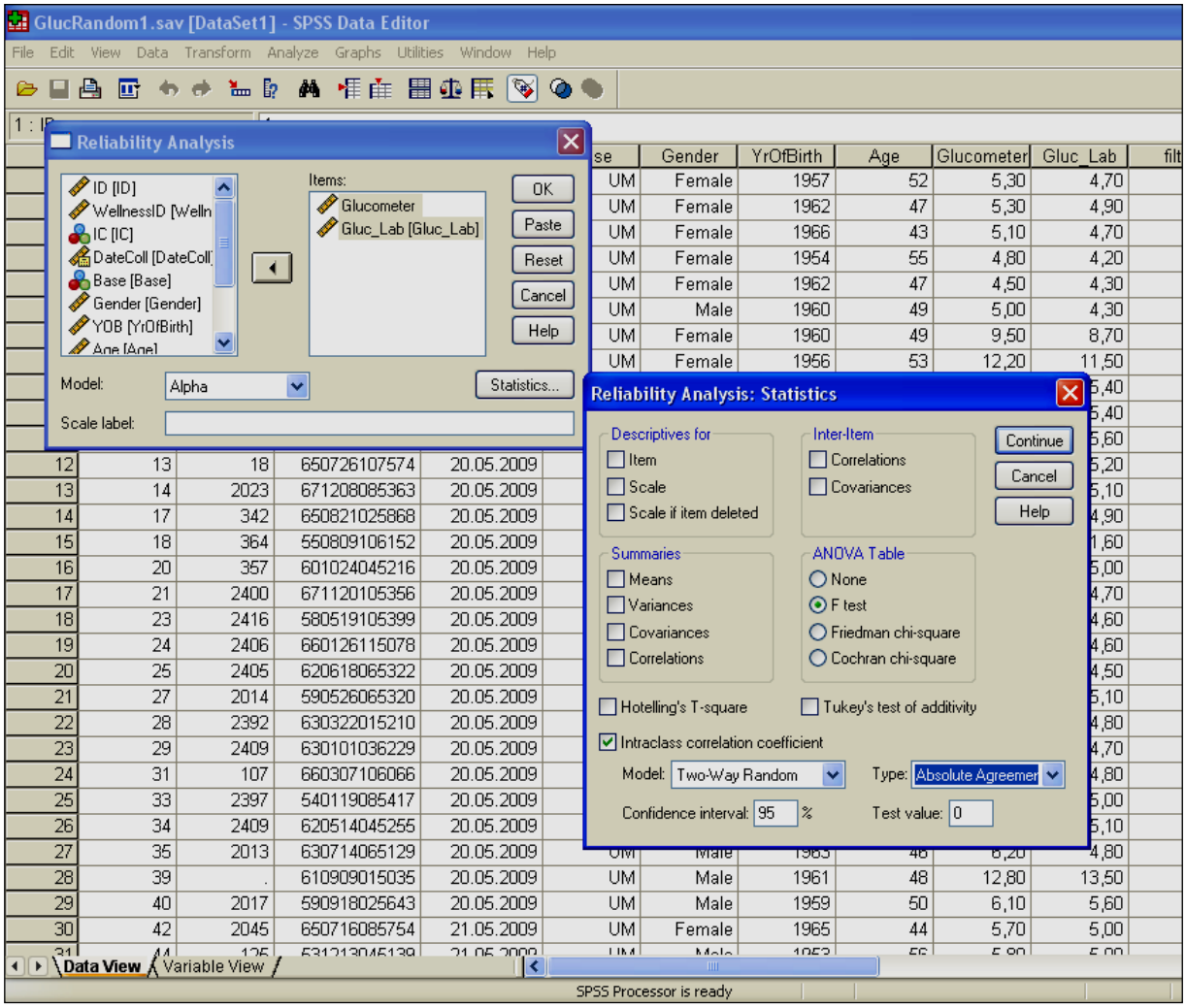


Figure 3.7: Analysis of ICC_A for agreement using SPSS (b)

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,954[b]	,840	,980	66,026	199	199	,000
Average Measures	,977	,912	,990	66,026	199	199	,000

Two-way random effects model where both people effects and measures effects are random.

a. Type A intraclass correlation coefficients using an absolute agreement definition.

b. The estimator is the same, whether the interaction effect is present or not.

Figure 3.8: Analysis of ICC_A for agreement: SPSS output

3.9.3 Method Used to Assess Reliability

3.9.3.1 Intra-class Correlation Coefficient for Reliability

The Intra-class Correlation Coefficient for consistency (ICC_C) analysis was performed using SPSS software, under reliability analysis with a consistency option. Since the purpose of the analysis was to test reliability, a two-way mixed model was chosen (Weir, 2005). The two-way mixed model assumes that each subject was assessed by the same raters, but these raters are the only ones of interest. Raters are considered as a fixed effect. In this study, the value of $ICC_C \geq 0.75$ was considered to be good reliability (Rosner, 2006).

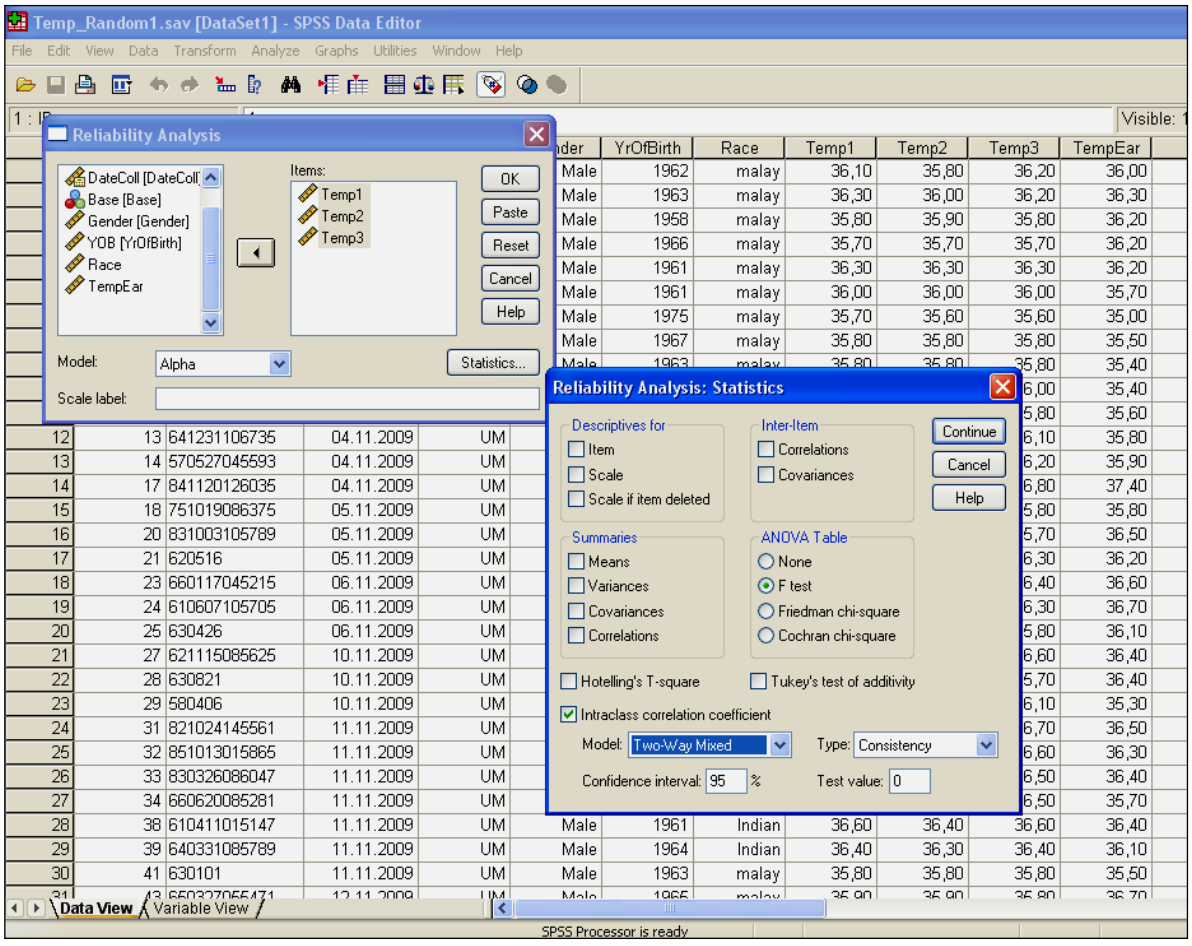


Figure 3.9: Analysis of ICC_C for reliability using SPSS

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,970[b]	,963	,977	98,828	199,0	398	,000
Average Measures	,990[c]	,987	,992	98,828	199,0	398	,000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Figure 3.10: Analysis of ICC_C for reliability: SPSS output

3.9.3.2 Intra-class Correlation Coefficient for Agreement

The analysis was performed using SPSS software and similar to as described in Section 3.9.2.3. However, three repeated readings from the same instrument were used. The value of $ICC_A \geq 0.75$ was considered to be good reliability (Rosner, 2006).

3.9.3.3 Bland-Altman Limits of Agreement

The analysis and interpretation of Bland-Altman method or reliability analysis were similar to the agreement analysis. However, it involved comparison of two repeated readings from the same instrument. The limits of agreement were compared with the tolerable clinical difference value (see Table 3.4). If the limits of agreement were less or equal to the tolerable clinical difference value, this means that there was an agreement between the two readings. Thus, the tested instrument was considered to be reliable.

Table 3.4: Tolerable clinical difference value set for reliability study

Variable	Tolerable clinical difference
Systolic Blood Pressure, SBP	±10mmHg (Pickering et al., 2005)
Diastolic Blood Pressure, DBP	±10mmHg (Pickering et al., 2005)
Peak Expiratory Flow Rate, PEFR	±40 l/min (Quanjer PH et al., 1997)
Heart rate, HR	±2bpm (Burke & Whelan, 1987)
Body temperature, Temp	$\pm 0.5^{0}$C (Robinson, Jou, & Spady, 2005)
Carbon monoxide level, CO level	±1ppm (*piCO+ Smokerlyzer User Manual*, 2006)

3.10 Data Entry and Cleaning

All primary data were recorded manually (written by hand) on a data sheet during the process of data collection. All raw data were then entered onto an SPSS 17.0 spreadsheet. Separate files were created for each variable to make the data more manageable. Information entered included: participant identity number, UM wellness identity number (where applicable), identity card number, date of data collection, base of data collection (UM or community), race, gender, year of birth and clinical values (e.g. glucometer value, laboratory value, etc). Age of participant (in years) was computed based on year of birth.

Double data entry was performed to ensure accuracy and minimize error during data entry. Two sets of files were created for the same variable, then those two files were compared using SPSS to see whether there were any differences. The general syntax for the function is shown in Figure 3.11. Any differences or missing values were double-checked with the original raw data set. The aim of data cleaning is to make sure that the data are free from as many errors as possible and data are of sufficient quality before they are used for analysis. Since the primary data collection was collected and

recorded carefully by the researcher, there were no missing data. After the files were finalised and saved as SPSS files, data were then transferred to other software (e.g. GraphPad Prism and Matlab) for further analysis.

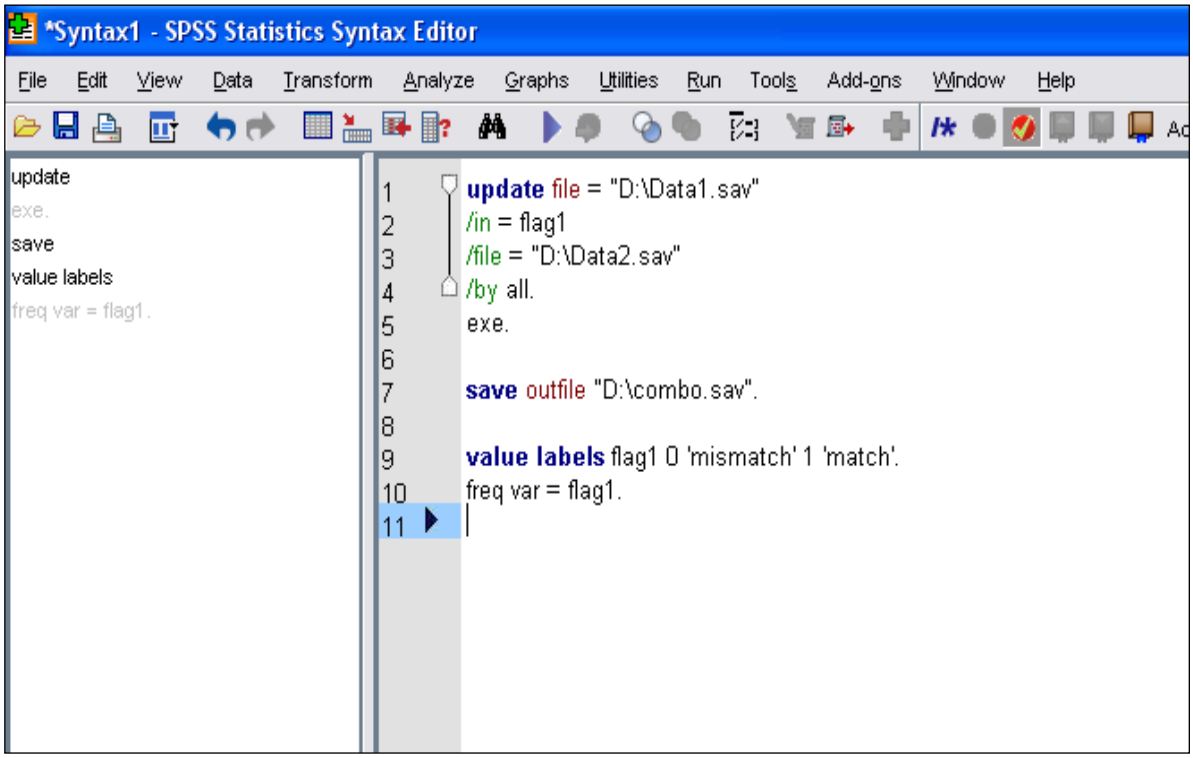


Figure 3.11: Syntax for data set comparison using SPSS

3.11 Data Analysis

3.11.1 Descriptive Analysis

A descriptive analysis was performed on the entire data set. The description of the sample for the study population was summarised according to the phases and site (UM and community setting) of data collection. This includes number of participants and their characteristics (age, race and gender). Summary measures (mean, median, standard deviation and range) for each variable were calculated and presented in a table. All data were assumed to be in a normal distribution (as according to the *Central Limit Theorem*), however the skewness and kurtosis of data set for each variable were also checked. All descriptive analyses were performed using SPSS 17.0 software.

3.11.2 Analysis of Agreement

The Bland-Altman method (Limits of Agreement, LoA) and the Intra-class Correlation Coefficient (ICC_A) are the two most commonly used methods to assess agreement in medicine found in Chapter 2. The purpose of this analysis is to assess the suitability of the comparison of slopes and y-intercepts anaysis and the proposed agreement model in comparison with those two methods (LoA and ICC_A). Analyses were performed using both original clinical data and simulated data.

3.11.2.1 Comparison of Statistical Methods: Clinical Data

a. Comparison of prediction of agreement

The first analysis was planned to test whether the proposed method is able to provide a similar prediction of bias and conclusion on the degree of agreement (i.e. good agreement or poor agreement), compared to the most commonly used methods to assess agreement. Analysis of agreement was conducted with a sample size of 300 (all the samples collected in this study) and was performed for the following variables: glucose level, systolic BP, diastolic BP, body temperature, peak expiratory flow rate and weight.

b. Consistency of prediction

For the second analysis, 10 sets of data with a sample of 200 were selected randomly from the total of 300 measurements (sampling with replacement) using SPSS software. The purpose of this analysis was to assess the consistency of the bias prediction and conclusion on the degree of agreement for all four methods. In this section, all results for the 10 sets of 200 random samples are compared in a table for each variable.

3.11.2.2 Comparison of Statistical Methods: Simulated Data

The purpose of data analysis in this section was to compare the performance of the comparison of slopes and y-intercepts analysis, agreement model, LoA and ICC_A according to the proportion of error in data, consistency of error (constant or inconsistent) and sample size. Analyses were performed for all the variables collected for agreement analysis (glucose level, systolic BP, diastolic BP, peak expiratory flow rate and body weight). The analysis was planned to assess whether these methods were

able to predict the simulated bias accordingly. Data sets with simulated bias were computed using the SPSS software.

1. Blood Glucose level: Laboratory value versus simulated value (±0.8mmol/l bias)
2. Systolic Blood Pressure (SBP): Manual sphygmomanometer SBP reading versus simulated value (±10mmHg bias)
3. Diastolic Blood Pressure: Manual sphygmomanometer DBP reading versus simulated value (±10mmHg bias)
4. Peak Expiratory Flow Rate (PEFR): Clement Clarke UK peak flow meter (first reading) versus simulated value (±40 l/min bias)
5. Weight: Digital weighing scale value versus simulated value (±0.5kg bias)

a. Constant systematic error

The purpose of this analysis is to assess the ability of each statistical method in predicting simulated constant bias in the data sets for all five variables.

a) Overestimation of value (positive error)

b) Underestimation of value (negative error)

b. Inconsistent error

The purpose of this analysis is to assess how each statistical method predicts the inconsistent error in the data sets. Data were selected randomly from data sets with simulated error according to the proportion of error as below:

a) 2/3 overestimate (positive error), 1/3 underestimate (negative error)

b) 1/3 overestimate (positive error), 2/3 underestimate (negative error)

c) 1/2 overestimate (positive error), 1/2 underestimate (negative error)

d) 1/3 overestimate (positive error), 1/3 underestimate (negative error), 1/3 agreement

c. Proportion of error

The aim of this analysis is to assess how the proportion of error in data influences the prediction of bias for each statistical method.

a) 1/3 of positive error, and 2/3 of agreement in data set
b) 1/2 of positive error, and 1/2 of agreement in data set
c) 2/3 of positive error, and 1/3 of agreement in data set
d) 3/4 of positive error, and 1/4 of agreement in data set

d. Sample size

The aim of this analysis is to see how sample size influences the prediction of bias for all three statistical methods. In this section, analyses of agreement using all four methods were performed for each variable for various sample sizes. For each variable, analysis of agreement using the Bland-Altman method, linear regression method and ICC_A were performed for sample sizes of 10 to 500. Sampling with replacement was performed. When a unit selected at random from the data set, it was returned back to the data set before the next sample was selected. Thus, whenever a unit is selected, the population remains the same size. This analysis was repeated ten times for each variable using randomly selected data sets from the original clinical data sets.

The outcome of the Bland-Altman analysis (estimated bias, upper limit of agreement and lower limit of agreement), slope, y-intercept, agreement model (predicted bias), and ICC_A were then plotted against the sample size to assess the pattern of the prediction. The standard deviation and standard error of the prediction were also plotted against the sample size. Analyses were performed using the Matlab programme version 7.8 (R2009a).

3.11.2.3 Extended Analysis of the Bland-Altman Method

As discussed in Chapter 2, one of the problems with the Bland-Altman analysis is a proportional bias. The main concern about the proportional bias is that this will result in artefactual bias or overestimation in the prediction. An approach using least-products regression to fit the regression line in the Bland-Altman plot has been claimed to eliminate the bias problem in the Bland-Altman analysis (J. Ludbrook, 2002). The purpose of this analysis is to determine whether the suggested method is able to overcome the overestimation of bias in the prediction.

Three variables (blood glucose level, body weight and systolic BP) with sample sizes of 300 for each variable were analysed in this section. The original data sets were compared with generated comparison data sets for all the variables. The comparison sets were generated with a random error of specific range of bias. Details of the generated bias will be explained in the next chapter (Chapter 4). For this analysis, data generation was performed using Matlab software version 7.8 (R2009a). The Bland-Altman analysis was performed for all the simulated data sets, for all three variables.

The range of predicted error was compared with the range of simulated error. The range of predicted error was determined based on the range between the minimum value of the lower limit of agreement and the maximum value of the upper limit of agreement (i.e. upper CI of upper limit of agreement minus lower CI of lower limit of agreement).

The presence of proportional bias was tested by testing the slope of the regression line fitted to the Bland-Altman plot. Proportional bias excluded when the slope did not significantly differ from zero. Analyses were performed using MedCalc statistical software version 12.1.3. Figure 3.12 shows an example of how the

proportional bias is excluded in the Bland-Altman plot. The regression analysis of the plot is:

y = 0.9322 - 0.01070 x

The slope = -0.01070 (95%CI -0.03533 to 0.01394), p = 0.3870

The slope of the regression line does not significantly differ from zero. Therefore the proportional bias is not present and should not be a cause of concern in the analysis.

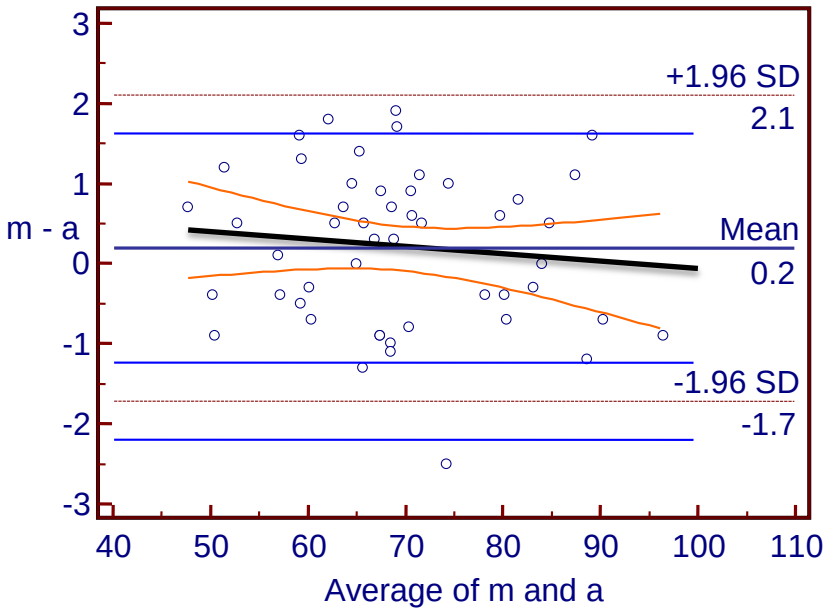


Figure 3.12**:** Regression line fitted to the Bland-Altman plot

Figure 3.13 shows the presence of proportional bias in the Bland-Altman plot. The regression analysis of the plot is:

$$y = -72.2265 + 0.7009 x$$

The slope = 0.7009 (95% CI 0.4671 to 0.9348), $p < 0.0001$

The slope of the regression line is significantly different from zero. Therefore proportional bias is present.

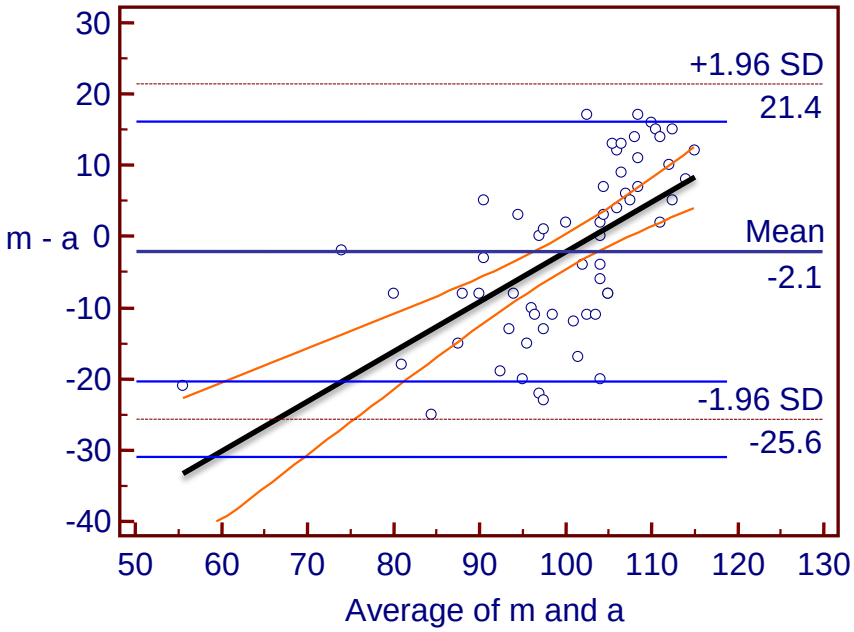


Figure 3.13: Example of the presence of proportional bias

3.11.3 Analysis of Reliability

The systematic review in Chapter 2 showed that the Intra-class Correlation Coefficient is the most popular method used to assess reliability in medicine. The Bland-Altman Limits of Agreement is also one of the most popular methods that have been used to assess reliability in medical research. The purpose of this analysis is to review the suitability of the Bland-Altman method (Limits of Agreement) for the measurement of precision (reliability), in comparison with the Intra-class Correlation Coefficient (both the ICC for consistency and ICC for agreement).

3.11.3.1 Comparison of Statistical Methods: Clinical Data

For the test of reliability, analyses using the Bland-Altman Limits of Agreement (LoA), the ICC for consistency (ICC_C) and the ICC for agreement (ICC_A) were compared. Analyses were performed for all variables collected for reliability analysis:

1. Systolic Blood Pressure (SBP): First, second and third readings of automatic SBP machine
2. Diastolic Blood Pressure (DBP): First, second and third readings of automatic DBP machine
3. Heart Rate (HR): First, second and third readings of automatic blood pressure machine
4. Body Temperature: First, second and third readings of non-contact infrared thermometer
5. Peak Expiratory Flow Rate (PEFR): First, second and third reading of Clement Clarke peak flow meter
6. Carbon monoxide (CO) level: First, second and third reading of carbon monoxide meter (piCO Smokerlyzer).

ICC analyses were performed with all three measurements for each variable. However, since the LoA was designed for the analysis of two repeated readings, the analyses of reliability were performed with the first and second measurements for this method. Although Bland and Altman suggested a method for the application of multiple measurements for the LoA analysis (Bland & Altman, 2007), it was only suitable for the analysis of agreement (as discussed in Chapter 2).

a. Measurement of reliability

The first analysis was planned to test whether the three methods (LoA, ICC_C and ICC_A) were able to provide a similar level of prediction or conclusion on reliability (e.g. good or poor reliability). For this section, the analysis of reliability was conducted with a sample of 300 for all variables.

b. Consistency of prediction

The purpose of this analysis was to assess the consistency of prediction on reliability for all three methods, and for all ten groups of data sets. For this section, 10 sets of data with a sample of 200 were selected randomly (sampling with replacement) from the total of 300 measurements using SPSS software. All results for the 10 groups of 200 random samples were compared for each variable.

c. Number of measurements

In this section, the results of reliability analysis for analysis of two sets of repeated measurements (first and second readings) were compared with analysis of three sets of repeated measurements (all three readings). The purpose of this analysis was to assess differences in the prediction of reliability with two and three sets of measurements. The LoA was excluded in the analysis.

3.11.3.2 Comparison of Statistical Methods: Simulated Data

Prediction of simulated data

In this analysis, five sets of simulated data were produced for each variable to represent five repeated measurements. The simulated data sets were designed to represent imprecise measurements of an instrument. This means that the five repeated measurements are not repeatable. The characteristics of the data sets were as follows:

Set 1 (first measurement): original clinical data

Set 2 (second measurement): 1/3 of data have constant positive error

Set 3 (third measurement): constant negative error

Set 4 (fourth measurement): 1/2 of data have constant positive error

Set 5 (fifth measurement): constant positive error

The results of analysis for ICC_C and ICC_A were compared. The LoA was excluded in the analysis.

3.11.3.3 Extended Analysis of the Intra-class Correlation Coefficient

In this section the results of reliability analysis for the analysis of two sets of repeated measurements (first and second readings) were compared with the analysis of three sets of repeated measurements (all three readings). The purpose of this analysis was to assess differences in the prediction of reliability with two and three sets of measurements.

For the clinical data, the analysis of reliability using ICC_C and ICC_A was performed for 10 sets of random samples of 200, for all six variables for both sets of measurements (two sets and three sets of measurements). The outcome of ICC for two

sets of measurements and three sets of measurements (with n=60) were tested using a paired t-test.

For the simulated data, the analysis of reliability using ICC_C and ICC_A was performed for two sets of random samples of 300, for all six variables for both sets of measurements (two sets and three sets of measurements). The outcome of ICC for two sets of measurements and three sets of measurements (with n=12) were tested using a paired t-test.

The paired t-test is used to test whether the difference between a pair of variables measured on each individual, on average, is zero (Kirkwood, 2000). In this study, a *p*-value of <0.05 was considered to be statistically significant. Statistical power was calculated for any non-significant result.

3.12 Summary of Chapter 3

This chapter details the major work done in this study, starting from the design of the study through to the analysis of data. The study population and study variables were described at the beginning of this chapter. The population of this study was selected from the institutional and community setting, to ensure the variability of the data as explained in Section 3.2. Details on the method of measuring the variables were also explained. All variables were measured according to the instruction in the instrument's manual or guidelines.

The data collection was divided into two phases. Phase I involved the UM Wellness Health-Screening programme and Community Health Screening in Teluk Gadong Kecil, Klang. Variables collected in this phase were glucose level, systolic BP, diastolic BP and heart rate. During the second phase (phase II), data were collected from the UM Wellness Quit Smoking Clinic and Community Health Screening in Mid Valley

Megamall. Variables collected in the second phase were CO level, PEFR, body weight and temperature. The data collection period was longer than expected due to various problems related to the UM Wellness screening schedule and slow response in the Quit Smoking Clinic. However, the response in the community was overwhelming, and helped to speed up the process of data collection.

Sample size in this study was estimated based on the linear regression method. A sample size of 200 was required, but 300 readings were collected for analysis purposes. Double data entry was performed to ensure the accuracy of the data.

The statistical software and statistical methods used in the analysis have been briefly described in this chapter. The method of assessing agreement using comparison of slopes and y-intercepts analysis was also explored here. An agreement model to estimate bias or error produced by an instrument was also proposed and explained in this chapter.

The analysis of data was divided into two sections (agreement analysis and reliability analysis). Statistical methods compared for agreement analysis were the comparison of slopes and y-intercepts analysis, agreement model, Bland-Altman Limits of Agreement and Intra-class Correlation Coefficient (ICC_A). Comparisons were made based on the ability of each method to predict simulated bias, the consistency of prediction and the effect of sample size on the prediction.

In the reliability analysis, the Intra-class Correlation Coefficient (both ICC_A for agreement and ICC_C for consistency) and the Bland-Altman Limits of Agreement were compared. The ability of each method to predict simulated bias, the consistency of prediction and the effect of the number of measurements were tested.

Most of the analyses were performed using SPSS and GraphPad Prism software. Matlab software was used mainly for data simulation (sampling with replacement

technique) to test the effect of sample size on the prediction. As a summary, the flow of this study is shown in Figure 3.14.

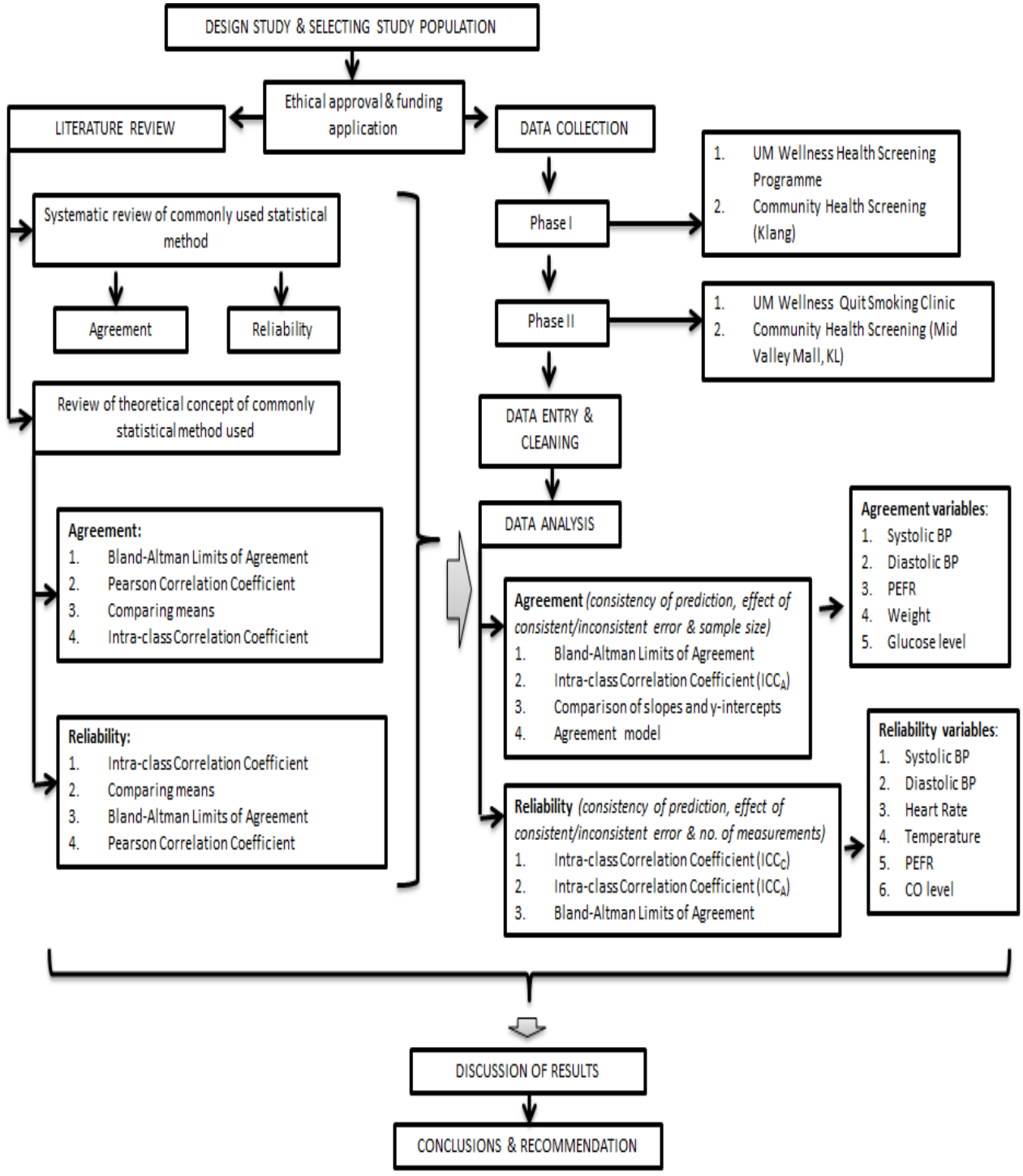


Figure 3.13: Study Flow

CHAPTER 4: RESULTS

4.1 Introduction

This chapter presents the results of statistical analysis of this project. Section 4.2 is a description of the sample and variables according to the phases of data collection. All variables will be described according to mean, median, range and standard deviation. The distributions of all variables are assumed to be normal (according to Central Limit Theorem); however the skewness and kurtosis of the data will also be presented.

The analysis begins with the analysis of agreement using clinical data in Section 4.3.1. Four methods of assessing agreement are compared; which are the comparison of slopes and y-intercepts analysis, agreement model, Bland-Altman Limits of Agreement (LoA) and Intra-class Correlation Coefficient for agreement (ICC_A). The aim is to see whether the proposed agreement model and comparison of slopes and y-intercepts analysis are able to provide similar prediction of bias and agreement (i.e. good agreement or poor agreement), with the other two methods. The next analysis is the analysis of agreement using simulated data (Section 4.3.2). The aim of this analysis is to compare the prediction of bias for each statistical method based on various sample size, range of variable, proportion of error in data, and consistency of error in data set. All statistical methods are tested to predict a known bias from generated data. In Section 4.3.3 the extended analysis of the Bland-Altman method is presented, which involve the testing of method to overcome proportional bias in the analysis.

Section 4.4 presents the analysis of reliability. This section compares different statistical methods that have been used to assess reliability which are the Bland-Altman method, Intra-Class Correlation Coefficient for consistency (ICC_C), and Intra-Class Correlation Coefficient for agreement (ICC_A). The data used are both clinical (Section

4.4.1) and simulated data (Section 4.4.2). Section 4.4.3 presents the extended analysis of the Intra-class Correlation Coefficient (for both ICC_A and ICC_C). This involves testing the differences between the prediction of ICCs with two and three repeated measurements.

The summary of the result is covered in Section 4.5. The results of all the analysis in this chapter provide information on the strengths and weaknesses of each statistical method and thus help in determining which statistical method is the best for assessing agreement and reliability.

4.2 Description of Samples and Variables

4.2.1 Phase I data collection

Four variables were collected during Phase I data collection (glucose, systolic BP, diastolic BP and pulse rate). A total of 300 samples were collected for each variable.

Data for glucose variable were collected from UM Wellness Health Screening program. The majority of participants were Malays (79.0%), followed by Indians (11.3%), Chinese (7.7%), and other races (2.0%). The mean age of participants was 48.6 years, and range between 39 to 67 years old. There were 170 (56.7%) male participants (56.7%) and 130 (43.3%) female participants.

Data for systolic BP, diastolic BP and pulse rate were collected from both the UM Wellness Health Screening Program (107 participants) and Community Health Screening Program in Klang (193 participants). A total of 300 samples were collected per variable. The majority of the participants were Malays (87.0%), followed by other races (6.0%), Indians (4.0%) and Chinese (3.0%). Fifty-four percent of the participants were female. Participants from the community in Klang had a wider range of age (min 12 to max 81 years old) compared to the participants from UM Wellness Health

Screening program (min 22 to max 63 years old). Table 4.1 summarises the description of sample in Phase I data collection.

Table 4.1: Description of Sample in Phase I

Variable collected	**Glucose**	**Systolic BP, Diastolic BP & Heart Rate**		
Sample population	UM Wellness Health Screening Program	UM Wellness Health Screening Program	Community health screening program Klang, Selangor	
Period of data collection	20/05/2009-12/06/2009	12/05/2009-27/08/2009	10/06/2009-24/07/2009	Total
Total participants	300	107	193	300
Age (years)	48.6 (39-67)	47.1 (22-63)	40.3 (12-81)	43.1 (12-81)
Gender				
▪ **Male**	170 (56.7%)	54 (50.5%)	84(43.5%)	138 (46.0%)
▪ **Female**	130 (43.3%)	53 (49.5%)	109 (56.5%)	162 (54.0%)
Race				
▪ **Malay**	237 (79.0%)	85 (79.4%)	176 (91.2%)	261 (87.0%)
▪ **Indian**	34 (11.3%)	11 (10.3%)	1 (0.5%)	12 (4.0%)
▪ **Chinese**	23 (7.7%)	9 (8.4%)	0	9 (3.0%)
▪ **Others**	6 (2.0%)	2 (1.9%)	16 (8.3%)	18 (6.0%)

4.2.2 Phase II data collection

Four variables were collected during Phase II data collection. These were body weight, body temperature (Temp), peak expiratory flow rate (PEFR), and carbon monoxide level (CO). A total of 300 samples were collected for each variable, with a total of 300 participants. Data were collected from UM Wellness Quit Smoking Clinic and UM Community Health Awareness Day Mid Valley Kuala Lumpur. The majority of the participants were from UM Wellness Quit Smoking Clinic (204 or 68.0%). Most of the participants in the Phase II data collection were male (85.3%), and all participants in the Quit Smoking Clinic were male. The mean age of participants was 37.3 years (with a range between 20 to 67 years old). Malays formed the biggest proportion of participants (90.0%), followed by Chinese (4.3%), Indian (3.7%) and other races (2.0%). Table 4.2 summarises the description of sample in Phase II data collection.

Table 4.2: Description of Sample in Phase II

Variable collected	**Weight, Temp, PEFR & CO**		
Sample population	UM Wellness Quit Smoking Clinic	UM Community Health Awareness Day Mid Valley, Kuala Lumpur	
Period of data collection	27/10/2009-12/06/2009	6/02/2010-7/02/2010	Total
Total participants	204 (68.0%)	96 (32.0%)	300
Age (years)	38.4 (20-58)	35.0 (21-67)	37.3 (20-67)
Gender			
▪ **Male**	204 (100%)	52 (54.2%)	256 (85.3%)
▪ **Female**	0	44 (45.8%)	44 (14.7%)
Race			
▪ **Malay**	195(95.6%)	75 (78.1%)	270 (90.0%)
▪ **Indian**	7 (3.4%)	4 (4.2%)	11 (3.7%)
▪ **Chinese**	0	13 (13.5%)	13 (4.3%)
▪ **Others**	2 (1.0%)	4(4.2%)	6 (2.0%)

4.2.3 Variables

Total of eight variables were collected at two phases for the purpose of data analysis of this study. For agreement analysis, variables collected were:

1. Glucose level: Glucometer reading versus laboratory value
2. Systolic BP: Manual sphygmomanometer BP reading versus automatic BP (first reading)
3. Diastolic BP: Manual sphygmomanometer BP reading versus automatic BP (first reading)
4. Peak Expiratory Flow Rate: Clement Clarke UK Peak flow meter (average reading) versus Respicare Peak flow meter
5. Weight: Analogue weighing scale versus digital weighing scale

For reliability analysis, variables collected are:

1. Systolic BP: First, second and third reading of automatic BP machine
2. Diastolic BP: First, second and third reading of automatic BP machine
3. Heart rate: First, second and third reading of automatic machine
4. Body temperature: First, second and third reading of infrared thermometer
5. Peak Expiratory Flow Rate: First, second and third reading of Clement Clarke Peak flow meter
6. Carbon monoxide level: First, second and third reading of carbon monoxide meter (Smokerlyzer)

4.2.3.1 Description of variables

Eight variables were collected for this study. Summary of measures for all eight variables are displayed in Tables 4.3 to 4.10. The mean value for blood glucose (Table 4.3) measured using the glucometer was 5.9mmol/l, slightly higher than the laboratory value (5.6mmol/l). The data collected covers quite good range of blood glucose level (both normal and abnormal values). The skewness and kurtosis of laboratory blood glucose and glucometer value were almost the same. These suggest that the shapes of distribution for both glucometer and laboratory values were similar (Table 4.3).

Table 4.3: Description of Blood Glucose sample

Summary of measures (N=300)	**Laboratory blood glucose value mmol/l**	**Glucometer mmol/l**
Range (minimum-maximum)	10.5 (3.7-14.2)	11.5 (3.8-15.3)
Mean	5.6	5.9
Median	5.0	5.3
Standard deviation	1.8	1.8
Skewness (standard error)	2.9 (0.14)	2.8 (0.14)
Kurtosis (standard error)	9.0 (0.28)	9.1(0.28)

Second variable collected in this study was systolic BP (SBP). Mean SBP reading from the manual sphygmomanometer was 123mmHg, whereas mean readings from the automatic BP machine were slightly lower: 121mmHg for the first reading and 119mmHg for the second and third readings. The range of SBP readings were higher in automatic BP machine compared to the manual sphygmomanometer. Overall, the range of SBP collected in this study covers quite good range of SBP values (normal and abnormal). The skewness and the kutosis of the manual SBP readings and the automatic readings were about the same. Thus, the shapes of distribution for manual and automatic SBP readings were almost similar (Table 4.4).

Table 4.4: Description of Systolic Blood Pressure

Summary of Measures (N=300)	**Manual reading mmHg**	**Automatic 1st reading mmHg**	**Automatic 2nd reading mmHg**	**Automatic 3rd reading mmHg**
Range (minimum-maximum)	108 (84-192)	135 (66-201)	113 (82-195)	111 (81-192)
Mean	123	121	119	119
Median	120	118	117	116
Standard deviation	18	21	19	18
Skewness (standard error)	0.64 (0.14)	0.81 (0.14)	0.86 (0.14)	0.71 (0.14)
Kurtosis (standard error)	0.96 (0.28)	0.97 (0.28)	1.16 (0.28)	1.00 (0.28)

The third variable was diastolic BP (DBP). The mean DBP reading from manual sphygmomanometer was 77mmHg. Whereas, mean readings from the automatic BP machine were 77mmHg for the first reading, 76mmHg for the second reading and 74mmHg for the third reading. The range of DBP values collected in this study covers quite good range of values (normal and abnormal) although, the ranges of DBP readings were higher in automatic BP machine compared to the manual sphygmomanometer (Table 4.5).

Table 4.5: Description of Diastolic Blood Pressure

Summary of Measures (N=300)	**Manual reading mmHg**	**Automatic 1st reading mmHg**	**Automatic 2nd reading mmHg**	**Automatic 3rd reading mmHg**
Range (minimum-maximum)	66 (44-110)	96 (38-134)	89 (42-131)	82 (43-125)
Mean	77	77	76	74
Median	78	77	75	75
Standard deviation	12	14	13	13
Skewness (standard error)	0.72 (0.14)	0.41 (0.14)	0.43 (0.14)	0.40 (0.14)
Kurtosis (standard error)	0.06 (0.28)	0.95 (0.28)	0.76 (0.28)	0.53 (0.28)

The shape of distribution for DBP reading from manual sphygmomanometer was slightly different from the shapes of distribution for readings from automatic BP machine (Figure 4.1). Automatic DBP values have higher peak of distribution and less skewness in comparison with manual DBP values (Figure 4.1)

The manual method requires auscultation of the blood pressure, whereas the OMRON HEM 907XL automatic BP machine depends on oscillometric method (Omron Instruction Manual, 2009). The auscultatory method relies on the observer to detect the audible sounds (Korotkoff sounds). The Korotkoff sounds ~~are~~ originate from

a combination of turbulent blood flow and oscillations of the arterial wall (Pickering et al, 2005). In this study, Phase V was used to identify the DBP as recommended by the guideline (NHF, 2009). Different mechanisms in detecting the DBP values might explain the differences in the distribution of readings between automatic DBP and manual DBP. Furthermore, disagreement exists as to whether Korotkoff phase IV or V correlates more accurately with the diastolic blood pressure (Pickering et al, 2005). Therefore it is uncertain whether the oscillometric method of detecting DBP would more correlate with Phase IV or V. Nonetheless, mean readings for manual DBP and automatic DBP (first reading) were the same.

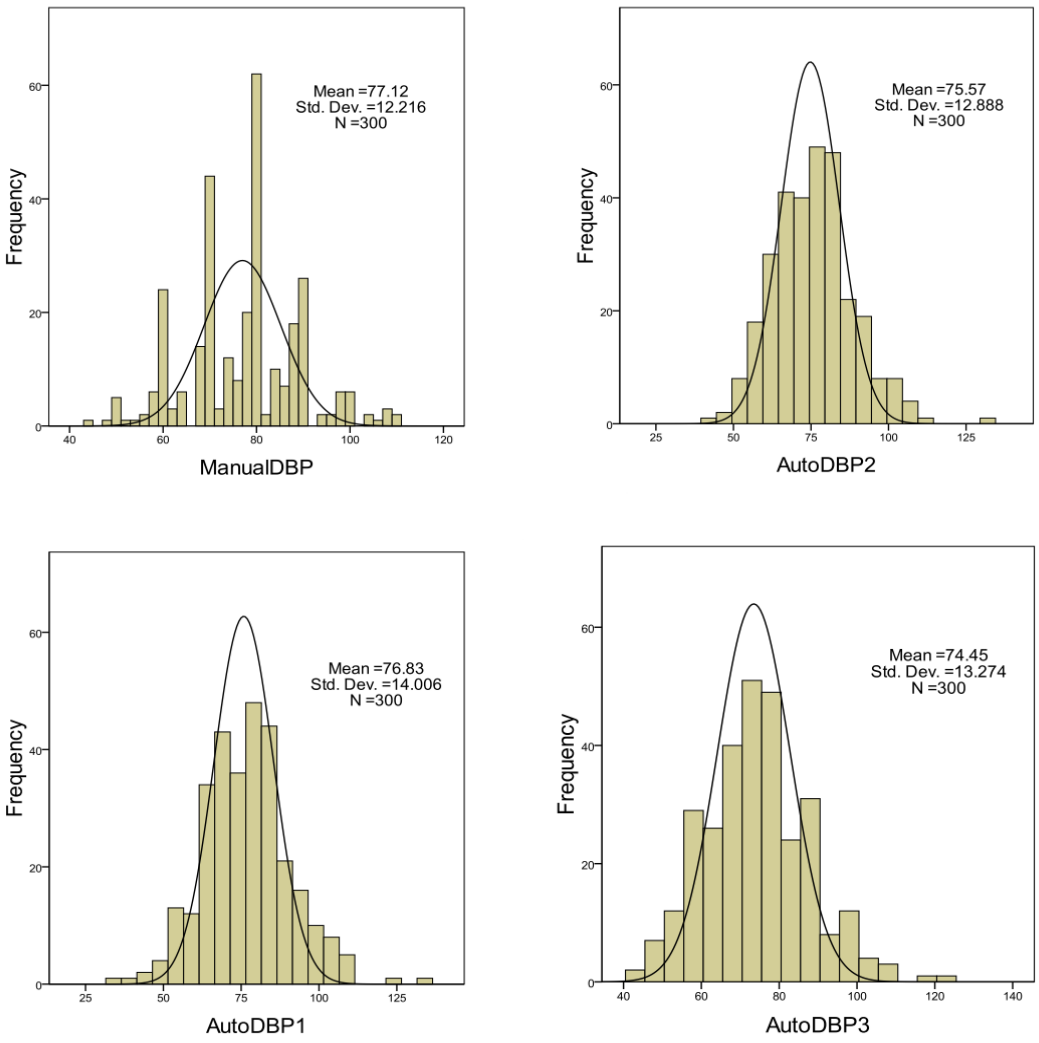


Figure 4.1: Comparison of shapes of distributions for the manual and automatic readings

The fourth variable was heart rate. The means and medians were the same (76 bpm) for all three readings. The range of values collected in this study covers mostly normal values of heart rate. The shapes of distribution of heart rate values were almost the same for the three readings (Table 4.6).

Table 4.6: Description of Heart Rate

Summary of Measures (N=300)	**Automatic 1st reading bpm**	**Automatic 2nd reading bpm**	**Automatic 3rd reading bpm**
Range (minimum-maximum)	53 (52-105)	48 (53-101)	50 (52-102)
Mean	76	76	76
Median	76	76	76
Standard deviation	11	11	11
Skewness (standard error)	0.09 (0.14)	0.09 (0.14)	0.10 (0.14)
Kurtosis (standard error)	-0.71 (0.28)	-0.77 (0.28)	-0.72 (0.28)

The next variable was body weight. The mean weights for both analogue and digital scale were 70kg. The range of weight for digital scale was slightly wider than analogue scale. However, both scales cover quite good range of body weight values (normal to abnormal). The shapes of distribution of values for both scales were about the same (Table 4.7).

Table 4.7: Description of Weight

Summary of Measures (N=300)	**Analog scale kg**	**Digital scale kg**
Range (minimum-maximum)	70 (40-110)	72.9 (38.5-111.4)
Mean	70	70.0
Median	70	69.5
Standard deviation	13	13
Skewness (standard error)	0.41 (0.14)	0.42 (0.14)
Kurtosis (standard error)	0.18 (0.28)	0.20 (0.28)

The mean temperatures for all three readings of non-contact infrared thermometer were the same (36.4^0C). The median temperatures for all three readings were also the same (36.5^0C). The range of values collected in this study covers mostly normal values of body temperature.

Table 4.8: Description of Temperature

Summary of Measures (N=300)	**Non-contact Infrared Thermometer 1st reading ^{0}C**	**Non-contact Infrared Thermometer 2nd reading ^{0}C**	**Non-contact Infrared Thermometer 3rd reading ^{0}C**
Range (minimum-maximum)	2.1 (35.0-37.1)	2.1 (35.0-37.1)	2.2 (35.0-37.2)
Mean	36.4	36.4	36.4
Median	36.5	36.5	36.5
Standard deviation	0.4	0.4	0.4
Skewness (standard error)	-0.73 (0.14)	-0.73 (0.14)	-0.70 (0.14)
Kurtosis (standard error)	0.39 (0.28)	0.23 (0.28)	0.23 (0.28)

Another variable collected in this study was Peak Expiratory Flow Rate (PEFR). The range of values collected in this study covers quite good range of PEFR values. The Respicare peak flow meter has higher mean and wider range of readings. The mean and median for all readings were almost the same except for the first reading of Clement Clarke Peak flow meter (Table 4.9).

Table 4.9: Description of Peak Expiratory Flow Rate

Summary of Measures (N=300)	**Respicare Peak flow meter l/min**	**Clement Clarke Peak flow meter Average reading l/min**	**Clement Clarke Peak flow meter 1st reading l/min**	**Clement Clarke Peak flow meter 2nd reading l/min**	**Clement Clarke Peak flow meter 3rd reading l/min**
Range (minimum-maximum)	530 (200-730)	464 (223-687)	470 (220-690)	500 (200-700)	460 (230-690)
Mean	471	453	441	455	464
Median	470	457	450	455	460
Standard deviation	86	81	84	86	85
Skewness (standard error)	-0.14 (0.14)	-0.079 (0.14)	-0.12 (0.14)	-0.07 (0.14)	0.71 (0.14)
Kurtosis (standard error)	0.23(0.28)	-0.062 (0.28)	-0.97 (0.28)	-0.04 (0.28)	1.00 (0.28)

The last variable collected in this study was carbon monoxide (CO) level. All three readings have the same mean (10ppm), median (9ppm), and standard deviation (7ppm). The ranges of all three readings were wide, covers both normal and abnormal value of CO level (Table 4.10).

Table 4.10: Description of Carbon Monoxide level

Summary of Measures (N=300)	**CO level 1st reading Ppm**	**CO level 2nd reading ppm**	**CO level 3rd reading ppm**
Range (minimum-maximum)	40 (1-41)	33 (1-34)	37 (1-38)
Mean	10	10	10
Median	9	9	9
Standard deviation	7	7	7
Skewness (standard error)	1.21 (0.14)	1.09 (0.14)	1.19 (0.14)
Kurtosis	1.71 (0.28)	0.72 (0.28)	1.21 (0.28)

4.3 Analysis of Agreement

The analysis under this section was planned to test whether the proposed agreement model and comparison of slopes and y-intercepts were able to provide better prediction of bias and conclusion on agreement, compared to the most commonly used methods (Bland-Altman Limits of Agreement and Intra-class Correlation Coefficient for agreement). The analysis was divided into two parts. The first part was the analysis of agreement using clinical data and the second part was analysis using simulated data. Analyses were performed using SPSS 17.0 and GraphPad Prism 5.02 software.

4.3.1 Comparison of statistical methods for Agreement analysis: Clinical data

4.3.1.1 Comparison of prediction of agreement

The comparison of slopes and y-intercepts analysis, agreement model, and two other most commonly used methods (LoA and ICC_A) were conducted with sample of 300 for each variable (blood glucose level, systolic BP, diastolic BP, peak expiratory flow rate and weight). The purpose of the first analysis is to compare the ability of each method in predicting agreement for each variable. Results for the analysis of each variable are summarised in Table 4.11 to Table 4.15.

Data used for the analysis were collected in clinical settings, and all instruments were validated by their manufacturer. The instruments should be in agreement with their standard instruments. The comparison of slopes and y-intercepts analysis suggests that there were agreement between SBP measured using manual sphygmomanometer and automatic BP machine, and also between the analogue weighing scale and digital weighing scale. However, no agreement was found for instruments that measure the

other three variables (blood glucose level, DBP and PEFR). The agreement model provides the same conclusion of agreement with ICC_A. Whereas the Bland-Altman LoA shows that there was no agreement for all medical instruments. The summaries of prediction of agreement for all four methods are displayed in Table 4.16.

Table 4.11: Comparison on prediction of agreement analysis for blood glucose level

Comparison of slopes and y-intercepts analysis	**Agreement Model**	**Bland-Altman LoA**	**ICC_A**
y_1 = 0.933x + 0.683 **r(Pearson) = 0.965** **r^2 = 0.931** **y_2 = x** Compare slopes: F-test: $p<0.0001$ Slopes – not equal	**Error= -0.067x + 0.683** Minimum glucose value = 3.7mmol/l Error = 0.44mmol/l Maximum glucose value = 14.2mmol/l Error = -0.27mmol/l Mean glucose value = 5.6mmol/l Mean error= -0.31mmol/l *Error between -0.31mmol/l to 0.44mmol/l Thus, errors produced by this instrument does not exceed clinically significant difference (i.e. will not have an effect clinically)	**Bias = 0.31mmol/l** **LoA = -0.62 to 1.24** Lower limit CI = (-0.7142 to -0.5293) Upper limit CI = (1.1500 to 1.3349) According to the LoA and its CI, errors produced by this instrument exceeded the clinically significant difference (i.e. will have an effect clinically)	**ICC_A = 0.950** CI (0.866 to 0.975)
Agreement – NO	**Agreement – YES**	**Agreement - NO**	**Agreement - YES**
Clinically significant difference for glucose = 0.8mmol/l (Essack et al., 2009)			

Table 4.12: Comparison on prediction of agreement analysis for SBP

Comparison of slopes and y-intercepts analysis	**Agreement Model**	**Bland-Altman LoA**	**ICC_A**
y_1 = 0.968x + 2.760 **r(Pearson) = 0.824** **r^2 = 0.679** **y_2 = x** Compare slopes: F-test = 0.675 *p* = 0.4115 Slopes – equal Compare intercept F-test = 2.683 *p* = 0.1019 Intercept – equal	**Error = -0.032x + 2.76** Minimum SBP value = 84mmHg Error = 0mmHg Maximum SBP value = 192mmHg Error = -3mmHg Mean SBP value = 123mmHg Mean error = -1mmHg *Error between -3mmHg to 0mmHg Thus, errors produced by this instrument does not exceed the clinically significant difference (i.e. will not have an effect clinically)	**Bias = -1 mmHg** **LoA = -25 to 22** Lower limit CI = (-27 to -22) Upper limit CI = (20 to 25) According to the LoA and its CI, errors produced by this instrument exceeded the clinically significant difference (i.e. will have an effect clinically)	**ICC_A = 0.812** CI (0.77 to 0.848)
Agreement – YES	**Agreement – YES**	**Agreement - NO**	**Agreement – YES**

Clinically significant difference for SBP = 10mmHg (Pickering et al., 2005)

Table 4.13: Comparison on prediction of agreement analysis for DBP

Comparison of slopes and y-intercepts analysis	**Agreement Model**	**Bland-Altman LoA**	**ICC_A**
y_1 = 0.8665x + 10.052 **r(Pearson) = 0.764** **r^2 = 0.584** **y_2 = x** Compare slopes: F-test = 9.9495 *p* = 0.00169 Slopes – not equal	**Error= -0.1335x+ 10.052** Minimum DBP value = 44mmHg Error = 4mmHg Maximum DBP value = 110mmHg Error = -5mmHg Mean DBP value = 77mmHg Mean error = 0mmHg *Error between -5mmHg to 4mmHg Thus, errors produced by this instrument does not exceed the clinically significant difference (i.e. will not have an effect clinically)	**Bias = -0.29 mmHg** **LoA = -18 to 18** Lower limit CI = (-20 to -17) Upper limit CI = (16 to 20) According to the LoA and its CI, errors produced by this instrument exceeded the clinically significant difference (i.e. will have an effect clinically)	**ICC_A = 0.759** CI (0.706 to 0.803)
Agreement – NO	**Agreement – YES**	**Agreement - NO**	**Agreement - YES**

Clinically significant difference for DBP = 10mmHg (Pickering et al., 2005)

Table 4.14: Comparison on prediction of agreement analysis for weight

Comparison of slopes and y-intercepts analysis	**Agreement Model**	**Bland-Altman LoA**	**ICC_A**
y_1 = 1.005x – 0.417 **r(Pearson) = 0.996** **r^2 = 0.992** **y_2 = x** Compare slopes: F-test = 0.8067 *p* = 0.3695 Slopes – equal Compare intercepts: F-test = 2.09159 *p* = 0.1486 Intercepts – equal	**Error = 0.005x -0.417** Minimum weight value = 38.5kg Error = -0.22kg Maximum weight value = 111.4kg Error = 0.14kg Mean weight value = 70kg Mean error = -0.07kg *Error between -0.22kg to 0.14kg Thus, errors produced by this instrument does not exceed the clinically significant difference (i.e. will not have an effect clinically)	**Bias = -0.09kg** **LoA = -2.39 to 2.20** Lower limit CI = (-2.61 to -2.16) Upper limit CI = (1.7 to 2.43) According to the LoA and its CI, errors produced by this instrument exceeded the clinically significant difference (i.e. will have an effect clinically)	**ICC_A = 0.996** CI (0.995 to 0.997)
Agreement – YES	**Agreement – YES**	**Agreement - NO**	**Agreement - YES**

Clinically significant difference for body weight = 0.5kg ("National Weights and Measures Laboratory," 2003)

Table 4.15: Comparison on prediction of agreement analysis for PEFR

Comparison of slopes and y-intercepts analysis	**Agreement Model**	**Bland-Altman LoA**	**ICC_A**
y_1 = 0.946x + 33.66 **r(Pearson) = 0.917** **r^2 = 0.841** **y_2 = x** Compare slopes: F-test = 2.15536 *p* = 0.1426 Slopes – equal Compare intercepts: F-test = 77.997 *p* <0.0001 Intercepts – not equal	**Error = -0.054x+ 33.66** Minimum PEFR value = 223 l/min. Error = 22 l/min Maximum PEFR value = 687 l/min Error = -3 l/min Mean PEFR value = 453 l/min Mean error = 9 l/min *Error between -3 l/min to 22 l/min Thus, errors produced by this instrument does not exceed the clinically significant difference (i.e. will not have an effect clinically)	**Bias = 17 l/min** **LoA = -50 to 85** Lower limit CI = (-56 to -43) Upper limit CI = (78 to 91) According to the LoA and its CI, errors produced by this instrument exceeded the clinically significant difference (i.e. will have an effect clinically)	**ICC_A = 0.896** CI (0.809 to 0.937)
Agreement – NO	**Agreement – YES**	**Agreement - NO**	**Agreement - YES**

Clinically significant difference for PEFR = 40l/min (Quanjer PH et al., 1997)

Table 4.16: Summary of prediction of agreement for all variables

Method	Glucose level	SBP	DBP	Weight	PEFR
1. Comparison of slopes and y-intercepts analysis	NO	YES	NO	YES	NO
2. Agreement model	YES	YES	YES	YES	YES
3. Bland-Altman LoA	NO	NO	NO	NO	NO
4. ICC_A	YES	YES	YES	YES	YES

NO - No agreement between instruments measuring the variable.
YES - Agreement between instruments measuring the variable.

4.3.1.2 Analysis of agreement: Consistency of prediction

For second analysis, 10 sets of data with sample of 200 were selected randomly (sampling with replacement) from the total of 300 samples for each variable. The purpose of this analysis is to compare the consistency of each statistical method in predicting agreement of instruments measuring each variable. Results for this analysis are summarised in Table 4.17 to Table 4.21.

The agreement model, Bland-Altman LoA, and ICC_A provide a consistent prediction of agreement for all ten sets of data for all variables. The comparison of slopes and intercepts analysis provides consistent prediction of agreement for DBP, weight and PEFR. The summaries of consistency of prediction of agreement for all four methods are displayed in Table 4.22.

Table 4.17: Comparison on consistency of agreement analysis for blood glucose level

Set	Comparison of slopes and y-intercepts analysis	Agreement Model	Bland-Altman LoA	ICC_A
1				
	y_1 = 0.937x + 0.677 **r = 0.9708** **y_2 = x** Slopes – not equal	**Error = -0.063x + 0.677** Glucose= 3.7to14.2mmol/l Error = 0.4 to -0.2mmol/l Mean glucose = 5.6mmo/l Mean error = 0.3mmol/l	**Bias = 0.3 mmol/l** LoA = -0.5 to 1.2 Lower limit CI = (-0.63 to -0.42) Upper limit CI = (1.07 to 1.28)	**ICC_A = 0.950** CI (0.866 to 0.975)
	Agreement – NO	**Agreement – YES**	**Agreement - NO**	**Agreement- YES**
2				
	y_1 = 0.924x + 0.752 **r = 0.9589** **y_2 = x** Slopes – not equal	**Error = -0.076x + 0.752** Glucose=3.7to 14.2mmol/l Error = 0.5 to -0.3mmol/l Mean glucose = 5.6 mmo/l Mean error = 0.3mmol/l	**Bias = 0.3 mmol/l** LoA = -0.6 to 1.2 Lower limit CI = (-0.63 to -0.42) Upper limit CI = (1.07 to 1.28)	**ICC_A = 0.940** CI (0.827 to 0.971)
	Agreement – NO	**Agreement – YES**	**Agreement - NO**	**Agreement- YES**
3				
	y_1 = 0.951x + 0.549 **r = 0.9724** **y_2 = x** Slopes – not equal	**Error = -0.049x + 0.549** Glucose=3.7to 14.2mmol/l Error = 0.4 to -0.2mmol/l Mean glucose = 5.6 mmo/l Mean error = 0.3mmol/l	**Bias = 0.3 mmol/l** LoA = -0.5 to 1.2 Lower limit CI = (-0.65 to -0.44) Upper limit CI = (1.08 to 1.29)	**ICC_A = 0.958** CI (0.861 to 0.981)
	Agreement – NO	**Agreement – YES**	**Agreement - NO**	**Agreement- YES**
4				
	y_1 = 0.933x + 0.699 **r = 0.9672** **y_2 = x** Slopes – not equal	**Error = -0.067x + 0.699** Glucose=3.7to 14.2mmol/l Error = 0.5 to -0.3mmol/l Mean glucose = 5.6 mmo/l Mean error = 0.3mmol/l	**Bias = 0.3 mmol/l** LoA = -0.5 to 1.2 Lower limit CI = (-0.65 to -0.44) Upper limit CI = (1.01 to 1.31)	**ICC_A = 0.950** CI (0.832 to 0.977)
	Agreement – NO	**Agreement – YES**	**Agreement - NO**	**Agreement- YES**
5				
	y_1 = 0.954x + 0.595 **r = 0.9680** **y_2 = x** Slopes – equal Intercepts – equal	**Error = -0.046x + 0.595** Glucose=3.7to 14.2mmol/l Error = 0.4 to -0.1mmol/l Mean glucose = 5.6 mmo/l Mean error = 0.3mmol/l	**Bias = 0.3 mmol/l** LoA = -0.6 to 1.3 Lower limit CI = (-0.68 to -0.46) Upper limit CI = (1.14 to 1.36)	**ICC_A = 0.952** CI (0.842 to 0.978)
	Agreement – YES	**Agreement – YES**	**Agreement - NO**	**Agreement- YES**
6				
	y_1 = 0.971x + 0.523 **r = 0.9595** **y_2 = x** Slopes – equal Intercepts – not equal	**Error = -0.029x + 0.523** Glucose=3.7to 14.2mmol/l Error = 0.4 to 0.1mmol/l Mean glucose = 5.6 mmo/l Mean error = 0.4mmol/l	**Bias = 0.4 mmol/l** LoA = -0.5 to 1.2 Lower limit CI = (-0.67 to -0.45) Upper limit CI = (1.20 to 1.43)	**ICC_A = 0.933** CI (0.725 to 0.973)
	Agreement – NO	**Agreement – YES**	**Agreement - NO**	**Agreement- YES**
7				
	y_1 = 0.929x + 0.706 **r = 0.9679** **y_2 = x** Slopes – not equal	**Error = -0.071x + 0.706** Glucose=3.7to 14.2mmol/l Error = -0.3 to 0.4mmol/l Mean glucose = 5.6 mmo/l Mean error = 0.3mmol/l	**Bias = 0.3 mmol/l** LoA = -0.5 to 1.2 Lower limit CI = (-0.65 to -0.44) Upper limit CI = (1.06 to 1.27)	**ICC_A = 0.952** CI (0.85 to 0.977)
	Agreement – NO	**Agreement – YES**	**Agreement - NO**	**Agreement- YES**

8				
	y_1 = 0.928x + 0.702	**Error = -0.072x + 0.702**	**Bias = 0.3 mmol/l**	**ICC_A = 0.956**
	r = 0.9688	Glucose=3.7to 14.2mmol/l	LoA = -0.6 to 1.2	CI (0.888 to 0.978)
	y_2 = x	Error = 0.4 to -0.3mmol/l	Lower limit CI	
	Slopes – not equal	Mean glucose = 5.6 mmo/l	= (-0.74 to -0.51)	
		Mean error = 0.3mmol/l	Upper limit CI	
			= (1.10 to 1.32)	
	Agreement – NO	**Agreement – YES**	**Agreement - NO**	**Agreement- YES**
9				
	y_1 = 0.927x + 0.695	**Error = -0.073x + 0.695**	**Bias = 0.3 mmol/l**	**ICC_A = 0.96**
	r = 0.9729	Glucose=3.7to 14.2mmol/l	LoA = -0.6 to 1.1	CI (0.887 to 0.980)
	y_2 = x	Error = 0.4 to -0.3mmol/l	Lower limit CI	
	Slopes – not equal	Mean glucose = 5.6 mmo/l	= (-0.66 to -0.45)	
		Mean error =	Upper limit CI	
			= (1.02 to 1.22)	
	Agreement – NO	**Agreement – YES**	**Agreement - NO**	**Agreement- YES**
10				
	y_1 = 0.924x + 0.747	**Error = -0.076x + 0.747**	**Bias = 0.3 mmol/l**	**ICC_A = 0.946**
	r = 0.9628	Glucose=3.7to 14.2mmol/l	LoA = -0.6 to 1.3	CI (0.848 to 0.974)
	y_2 = x	Error = 0.5 to -0.3mmol/l	Lower limit CI	
	Slopes – not equal	Mean glucose = 5.6 mmo/l	= (-0.74 to -0.51)	
		Mean error = 0.3mmol/l	Upper limit CI	
			= (1.16 to 1.39)	
	Agreement – NO	**Agreement – YES**	**Agreement - NO**	**Agreement- YES**

Clinically significant difference for glucose = 0.8mmol/l (Essack et al., 2009)

Table 4.18: Comparison on consistency of agreement analysis for SBP

Set	Comparison of slopes and y-intercepts analysis	Agreement Model	Bland-Altman LoA	ICC_A
1				
	y_1 = 0.990x + 0.451 **r = 0.8582** **y_2 = x** Slopes – equal Intercepts – equal	**Error = -0.01x + 0.451** SBP = 84 to 192mmHg Error = -0.4 to -1mmHg Mean SBP = 123mmHg Mean error = -1 mmHg	**Bias = -1 mmHg** LoA = -21 to 20 Lower limit CI = (-24 to -19) Upper limit CI = (17 to 22)	**ICC_A = 0.849** CI (0.806 to 0.884)
	Agreement – YES	**Agreement – YES**	**Agreement - NO**	**Agreement – YES**
2				
	y_1 = 0.999x - 0.548 **r = 0.8582** **y_2 = x** Slopes – equal Intercepts – equal	**Error = -0.001x – 0.548** SBP = 84 to 192mmHg Error = -0.6 to -0.7mmHg Mean SBP = 123mmHg Mean error = -1 mmHg	**Bias = -1 mmHg** LoA = -23 to 21 Lower limit CI = (-25 to -20) Upper limit CI = (19 to 24)	**ICC_A = 0.849** CI (0.805 to 0.883)
	Agreement – YES	**Agreement – YES**	**Agreement - NO**	**Agreement - YES**
3				
	y_1 = 1.013x – 2.728 **r = 0.8280** **y_2 = x** Slopes – not equal	**Error = -0.013x – 2.728** SBP = 84 to 192mmHg Error = -4 to -5mmHg Mean SBP = 123mmHg Mean error = -1 mmHg	**Bias = -1 mmHg** LoA = -52 to 50 Lower limit CI = (-26 to -21) Upper limit CI = (19 to 24)	**ICC_A = 0.810** CI (0.757 to 0.853)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
4				
	y_1 = 0.939x + 5.604 **r = 0.8176** **y_2 = x** Slopes – equal Intercepts – not equal	**Error = -0.061x + 5.604** SBP = 84 to 192mmHg Error = 0.5 to -6mmHg Mean SBP = 123mmHg Mean error = -2 mmHg	**Bias = -2 mmHg** LoA = -24 to 21 Lower limit CI = (-27 to -22) Upper limit CI = (18 to 24)	**ICC_A = 0.950** CI (0.832 to 0.977)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
5				
	y_1 = 0.987x + 0.565 **r = 0.8644** **y_2 = x** Slopes – equal Intercepts - equal	**Error = -0.013x + 0.565** SBP = 84 to 192mmHg Error = -0.5 to -2mmHg Mean SBP = 123mmHg Mean error = -1 mmHg	**Bias = -1 mmHg** LoA = -22 to 20 Lower limit CI = (-24 to -19) Upper limit CI = (17 to 22)	**ICC_A = 0.856** CI (0.814 to 0.889)
	Agreement – YES	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
6				
	y_1 = 0.929x + 6.875 **r = 0.8266** **y_2 = x** Slopes – equal Intercepts – not equal	**Error = -0.071x + 6.875** SBP = 84 to 192mmHg Error = 1 to -7 mmHg Mean SBP = 123mmHg Mean error = -2 mmHg	**Bias = -2 mmHg** LoA = -24 to 21 Lower limit CI = (-27 to -22) Upper limit CI = (18 to 23)	**ICC_A = 0.818** CI (0.766 to 0.860)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
7				
	y_1 = 0.963x – 3.149 **r = 0.8213** **y_2 = x** Slopes – equal Intercepts - equal	**Error = -0.037x – 3.149** SBP = 84 to 192mmHg Error = --6 to 10mmHg Mean SBP = 123mmHg Mean error = -1 mmHg	**Bias = -1 mmHg** LoA = -24 to 21 Lower limit CI = (-27 to -21) Upper limit CI = (18 to 24)	**ICC_A = 0.810** CI (0.756 to 0.853)
	Agreement – YES	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**

8				
	y_1 = 1.001x – 1.709 **r = 0.8370** **y_2 = x** Slopes – equal Intercepts - equal	**Error = 0.001x – 1.709** SBP = 84 to 192mmHg Error = -1.6 to -1.5mmHg Mean SBP = 123mmHg Mean error = -2 mmHg	**Bias = -2 mmHg** LoA = -25 to 22 Lower limit CI = (-28 to -23) Upper limit CI = (19 to 25)	**ICC_A = 0.822** CI (0.771 to 0.863)
	Agreement – YES	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
9				
	y_1 = 0.979x + 1.512 **r = 0.8069** **y_2 = x** Slopes – equal Intercepts - equal	**Error = -0.021x + 1.512** SBP = 84 to 192mmHg Error = -0.3 to -3mmHg Mean SBP = 123mmHg Mean error = -1 mmHg	**Bias = -1 mmHg** LoA = -25 to 23 Lower limit CI = (-28 to -22) Upper limit CI = (20 to 26)	**ICC_A = 0.792** CI (0.734 to 0.838)
	Agreement – YES	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
10				
	y_1 = 0.960x + 4.005 **r = 0.8307** **y_2 = x** Slopes – equal Intercepts - equal	**Error = -0.04x + 4.005** SBP = 84 to 192mmHg Error = 0.6 to -4mmHg Mean SBP = 123mmHg Mean error = -1 mmHg	**Bias = -1 mmHg** LoA = -24 to 23 Lower limit CI = (-27 to -22) Upper limit CI = (20 to 24)	**ICC_A = 0.822** CI (0.771 to 0.862)
	Agreement – YES	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**

Clinically significant difference for SBP = 10mmHg (Pickering et al., 2005)

Table 4.19: Comparison on consistency of agreement analysis for DBP

Set	Comparison of slopes and y-intercepts analysis	Analysis of Error	Bland-Altman LoA	ICC_A
1				
	y_1 = 0.819x +13.593 **r = 0.7574** **y_2 = x** Slopes – not equal	**Error = -0.181x + 13.593** DBP = 44 to 110 mmHg Error = 7 to -6mmHg Mean DBP = 77mmHg Mean error = -0.3 mmHg	**Bias = -0.4 mmHg** LoA = -18 to 17 Lower limit CI = (-20 to -16) Upper limit CI = (15 to 19)	**ICC_A = 0.756** CI (0.689 to 0.809)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
2				
	y_1 = 0.839x + 11.828 **r = 0.7493** **y_2 = x** Slopes – not equal	**Error = -0.161x + 11.828** DBP = 44 to 110 mmHg Error = 5 to -6mmHg Mean DBP = 77mmHg Mean error = -0.6 mmHg	**Bias = -0.5 mmHg** LoA = -18 to 19 Lower limit CI = (-21 to -16) Upper limit CI = (17 to 22)	**ICC_A = 0.745** CI (0.676 to 0.801)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
3				
	y_1 = 0.853x + 11.277 **r = 0.7593** **y_2 = x** Slopes – not equal	**Error = -0.147x + 11.277** DBP = 44 to 110 mmHg Error = 5 to -5mmHg Mean DBP = 77mmHg Mean error = -0.04 mmHg	**Bias = -0.1 mmHg** LoA = -17 to 17 Lower limit CI = (-19 to -15) Upper limit CI = (15 to 19)	**ICC_A = 0.754** CI (0.687 to 0.808)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
4				
	y_1 = 0.892x + 8.071 **r = 0.7770** **y_2 = x** Slopes – not equal	**Error = -0.108x + 8.071** DBP = 44 to 110 mmHg Error = 3 to -4mmHg Mean DBP = 77mmHg Mean error = -0.2 mmHg	**Bias = -0.2 mmHg** LoA = -18 to 18 Lower limit CI = (-20 to -16) Upper limit CI = (15 to 20)	**ICC_A = 0.77** CI (0.707 to 0.821)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
5				
	y_1 = 0.951x + 3.595 **r = 0.7812** **y_2 = x** Slopes – equal Intercepts - equal	**Error = -0.049x + 3.595** DBP = 44 to 110 mmHg Error = 1 to -2 mmHg Mean DBP = 77mmHg Mean error = -0.2 mmHg	**Bias = -0.3 mmHg** LoA = -17 to 17 Lower limit CI = (-19 to -15) Upper limit CI = (15 to 20)	**ICC_A = 0.767** CI (0.703 to 0.819)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
6				
	y_1 = 0.933x + 4.857 **r = 0.7882** **y_2 = x** Slopes – equal Intercepts – equal	**Error = -0.067x + 4.857** DBP = 44 to 110 mmHg Error = 2 to -3mmHg Mean DBP = 77mmHg Mean error = -0.3 mmHg	**Bias = -0.4 mmHg** LoA = -18 to 17 Lower limit CI = (-20 to -16) Upper limit CI = (15 to 20)	**ICC_A = 0.778** CI (0.717 to 0.827)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
7				
	y_1 = 0.812x + 14.361 **r = 0.7821** **y_2 = x** Slopes – not equal	**Error = -0.188x + 14.361** DBP = 44 to 110 mmHg Error = 6 to -6mmHg Mean DBP = 77mmHg Mean error = -0.1 mmHg	**Bias = -0.1 mmHg** LoA = -17 to 17 Lower limit CI = (-20 to -15) Upper limit CI = (15 to 19)	**ICC_A = 0.749** CI (0.681 to 0.804)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**

8				
	y_1 = 0.879x + 9.616 **r = 0.7463** **y_2 = x** Slopes – not equal	**Error = -0.121x + 9.616** DBP = 44 to 110 mmHg Error = 4 to -4mmHg Mean DBP = 77mmHg Mean error = 0.3 mmHg	**Bias = 0.3 mmHg** LoA = -18 to 19 Lower limit CI = (-20 to -16) Upper limit CI = (17 to 21)	**ICC_A = 0.737** CI (0.669 to 0.795)
	Agreement – NO	**Agreement - YES**	**Agreement – NO**	**Agreement - YES**
9				
	y_1 = 0.873x + 10.269 **r = 0.7706** **y_2 = x** Slopes –not equal	**Error = -0.127x + 10.269** DBP = 44 to 110 mmHg Error = 5 to -4mmHg Mean DBP = 77mmHg Mean error = 0.4 mmHg	**Bias = 0.4 mmHg** LoA = -17 to 18 Lower limit CI = (-19 to -15) Upper limit CI = (15 to 20)	**ICC_A = 0.769** CI (0.701 to 0.817)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**
10				
	y_1 = 0.857x + 10.531 **r = 0.7465** **y_2 = x** Slopes – not equal	**Error = -0.143x + 10.531** DBP = 44 to 110 mmHg Error = 4 to -5 mmHg Mean DBP = 77mmHg Mean error = -0.5 mmHg	**Bias = -1 mmHg** LoA = -19 to 18 Lower limit CI = (-21 to -17) Upper limit CI = (15 to 20)	**ICC_A = 0.74** CI (0.67 to 0.797)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**

Clinically significant difference for DBP = 10mmHg (Pickering et al., 2005)

Table 4.20: Comparison on consistency of agreement analysis for weight

Set	Comparison of slopes and y-intercepts analysis	Analysis of Error	Bland-Altman LoA	ICC_A
1				
	y_1 = 1.003x – 0.298 **r = 0.9947** **y_2 = x** Slopes – equal Intercepts – equal	**Error = 0.003x – 0.298** Weight = 38.5 to 111.4kg Error = -0.2 to 0.04kg Mean weight = 70kg Mean error = -0.1kg	**Bias = -0.1kg** LoA = -2.3 to 2.1 Lower limit CI = (-2.51 to -2.08) Upper limit CI = (1.84 to 2.27)	**ICC_A = 0.995** CI (0.993 to 0.996)
	Agreement – YES	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**
2				
	y_1 = 1.000x -0.032 **r = 0.9969** **y_2 = x** Slopes – equal Intercepts – equal	**Error = 0.000x - 0.032** Weight = 38.5 to 111.4kg Error = -0.03kg Mean weight = 70kg Mean error = -0.03kg	**Bias = -0.1kg** LoA = -2.0 to 1.9 Lower limit CI = (-2.25 to -1.78) Upper limit CI = (1.65 to 2.13)	**ICC_A = 0.997** CI (0.996 to 0.998)
	Agreement – YES	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**
3				
	y_1 = 1.003x – 0.210 **r = 0.9969** **y_2 = x** Slopes – equal Intercepts – equal	**Error = 0.003x – 0.210** Weight = 38.5 to 111.4kg Error = -0.1 to 0.1kg Mean weight = 70kg Mean error = 0.0kg	**Bias = -0.03kg** LoA = -2.0 to 2.0 Lower limit CI = (-2.29 to -1.80) Upper limit CI = (1.75 to 2.24)	**ICC_A = 0.997** CI (0.996 to 0.998)
	Agreement – YES	**Agreement - YES**	**Agreement – NO**	**Agreement – YES**
4				
	y_1 = 1.002x – 0.266 **r = 0.9952** **y_2 = x** Slopes – equal Intercepts – equal	**Error = 0.002x – 0.266** Weight = 38.5 to 111.4kg Error = -0.2 to -0.04kg Mean weight = 70kg Mean error = -0.1kg	**Bias = -0.1kg** LoA = -2.5 to 2.3 Lower limit CI = (-2.82 to -2.24) Upper limit CI = (2.01 to 2.60)	**ICC_A = 0.995** CI (0.994 to 0.996)
	Agreement – YES	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**
5				
	y_1 = 1.009x -0.698 **r = 0.9946** **y_2 = x** Slopes – equal Intercepts - equal	**Error = 0.009x -0.698** Weight = 38.5 to 111.4kg Error = -0.4 to 0.3kg Mean weight = 70kg Mean error = -0.1kg	**Bias = -0.1kg** LoA = -2.6 to 2.4 Lower limit CI = (-2.87 to -2.26) Upper limit CI = (2.09 to 2.69)	**ICC_A = 0.994** CI (0.993 to 0.996)
	Agreement – YES	**Agreement - YES**	**Agreement – NO**	**Agreement – YES**
6				
	y_1 = 1.004x – 0.398 **r = 0.9955** **y_2 = x** Slopes – equal Intercepts – equal	**Error = 0.004x – 0.398** Weight = 38.5 to 111.4kg Error = -0.2 to 0.05kg Mean weight = 70kg Mean error = -0.1kg	**Bias = -0.1kg** LoA = -2.6 to 2.3 Lower limit CI = (-2.88 to -2.28) Upper limit CI = (2.04 to 2.63)	**ICC_A = 0.995** CI (0.994 to 0.997)
	Agreement – YES	**Agreement - YES**	**Agreement – NO**	**Agreement – YES**
7				
	y_1 = 1.004x – 0.314 **r = 0.9968** **y_2 = x** Slopes – equal Intercepts – equal	**Error = 0.004x – 0.314** Weight = 38.5 to 111.4kg Error = -0.2 to 0.1kg Mean weight = 70kg Mean error = -0.03kg	**Bias = -0.03kg** LoA = -2.0 to 2.0 Lower limit CI = (-2.27 to -1.79) Upper limit CI = (1.73 to 2.22)	**ICC_A = 0.997** CI (0.996 to 0.998)
	Agreement – YES	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**

8				
	y_1 = 1.006x – 0.444	**Error = 0.006x – 0.444**	**Bias = -0.1kg**	**ICC_A = 0.996**
	r = 0.9956	Weight = 38.5 to 111.4kg	LoA = -2.6 to 2.4	CI (0.994 to 0.997)
	y_2 = x	Error = -0.2 to 0.2kg	Lower limit CI	
	Slopes – equal	Mean weight = 70kg	= (-2.86 to -2.25)	
	Intercepts – equal	Mean error = -0.02kg	Upper limit CI	
			= (2.14 to 2.74)	
	Agreement – YES	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**
9				
	y_1 = 1.001x – 0.166	**Error = 0.001x – 0.166**	**Bias = -0.1kg**	**ICC_A = 0.996**
	r = 0.9960	Weight = 38.5 to 111.4kg	LoA = -2.5 to 2.2	CI (0.995 to 0.997)
	y_2 = x	Error = -0.1 to -0.05kg	Lower limit CI	
	Slopes – equal	Mean weight = 70kg	= (-2.76 to -2.19)	
	Intercepts – equal	Mean error = -0.1kg	Upper limit CI	
			= (1.96 to 2.53)	
	Agreement – YES	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**
10				
	y_1 = 1.00x – 0.538	**Error = 0.006x – 0.538**	**Bias = -0.1kg**	**ICC_A = 0.997**
	r = 0.9972	Weight = 38.5 to 111.4kg	LoA = -2.1 to 1.9	CI (0.996 to 0.998)
	y_2 = x	Error = -0.3 to 0.1kg	Lower limit CI	
	Slopes – equal	Mean weight = 70kg	= (-2.34 to -1.86)	
	Intercepts – equal	Mean error = -0.1kg	Upper limit CI	
			= (1.61 to 2.09)	
	Agreement – YES	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**

Clinically significant difference for body weight = 0.5kg ("National Weights and Measures Laboratory," 2003)

Table 4.21: Comparison on consistency of agreement analysis for PEFR

Set	Comparison of slopes and y-intercepts analysis	Analysis of Error	Bland-Altman LoA	ICC_A
1				
	y_1 = 1.019x +10.545 **r = 0.9305** **y_2 = x** Slopes – equal Intercepts – not equal	**Error = 0.019x + 10.545** PEFR = 223 to 687 l/min Error = 15 to 24 l/min Mean PEFR= 453 l/min Mean error = 19 l/min	**Bias = 19 l/min** LoA = -45 to 83 Lower limit CI = (-53 to -37) Upper limit CI = (75 to 91)	**ICC_A = 0.905** CI (0.794 to 0.948)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
2				
	y_1 = 0.954x + 40.286 **r = 0.9081** **y_2 = x** Slopes – equal Intercepts – not equal	**Error = -0.046x + 40.286** PEFR = 223 to 687 l/min Error = 24 to 8 l/min Mean PEFR= 453 l/min Mean error = 19 l/min	**Bias = 20 l/min** LoA = -50 to 89 Lower limit CI = (-59 to -42) Upper limit CI = (81 to 98)	**ICC_A = 0.883** CI (0.766 to 0.932)
	Agreement - NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
3				
	y_1 = 0.966x + 31.488 **r = 0.9153** **y_2 = x** Slopes – equal Intercepts – not equal	**Error = -0.034x + 31.488** PEFR = 223 to 687 l/min Error = 24 to 8 l/min Mean PEFR= 453 l/min Mean error = 16 l/min	**Bias = 16 l/min** LoA = -52 to 85 Lower limit CI = (-60 to -44) Upper limit CI = (76 to 93)	**ICC_A = 0.898** CI (0.822 to 0.936)
	Agreement - NO	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**
4				
	y_1 = 0.967x + 31.959 **r = 0.9280** **y_2 = x** Slopes – equal Intercepts – not equal	**Error = -0.033x + 31.959** PEFR = 223 to 687 l/min Error = 25 to 9 l/min Mean PEFR= 453 l/min Mean error = 17 l/min	**Bias = 17 l/min** LoA = -48 to 81 Lower limit CI = (-55 to -40) Upper limit CI = (73 to 89)	**ICC_A = 0.910** CI (0.828 to 0.947)
	Agreement - NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
5				
	y_1 = 0.978x + 26.598 **r = 0.9210** **y_2 = x** Slopes – equal Intercepts – not equal	**Error = -0.022x + 26.598** PEFR = 223 to 687 l/min Error = 22 to 11 l/min Mean PEFR= 453 l/min Mean error = 17 l/min	**Bias = 17 l/min** LoA = -50 to 83 Lower limit CI = (-58 to -42) Upper limit CI = (75 to 92)	**ICC_A = 0.902** CI (0.823 to 0.940)
	Agreement - NO	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**
6				
	y_1 = 1.002x + 15.041 **r = 0.9404** **y_2 = x** Slopes – equal Intercepts – not equal	**Error = 0.002x + 15.041** PEFR = 223 to 687 l/min Error = 15 to 16 l/min Mean PEFR= 453 l/min Mean error = 16 l/min	**Bias = 16 l/min** LoA = -43 to 75 Lower limit CI = (-51 to -36) Upper limit CI = (68 to 83)	**ICC_A = 0.923** CI (0.846 to 0.955)
	Agreement - NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
7				
	y_1 = 0.978x + 25.268 **r = 0.9164** **y_2 = x** Slopes – equal Intercepts – not equal	**Error = -0.022x + 25.268** PEFR = 223 to 687 l/min Error = 20 to 10 l/min Mean PEFR= 453 l/min Mean error = 15 l/min	**Bias = 15 l/min** LoA = -51 to 82 Lower limit CI = (-59 to -43) Upper limit CI = (74 to 90)	**ICC_A = 0.899** CI (0.827 to 0.936)
	Agreement - NO	**Agreement - YES**	**Agreement - NO**	**Agreement – YES**

8				
	y_1 = 0.956x + 37.317	**Error = -0.044x + 37.317**	**Bias = 17 l/min**	**ICC_A = 0.885**
	r = 0.9066	PEFR = 223 to 687 l/min	LoA = -52 to 87	CI (0.794 to 0.930)
	y_2 = x	Error = 28 to 7 l/min	Lower limit CI	
	Slopes – equal	Mean PEFR= 453 l/min	= (-60 to -44)	
	Intercepts – not equal	Mean error = 17 l/min	Upper limit CI	
			= (78 to 95)	
	Agreement - NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
9				
	y_1 = 0.949x + 41.076	**Error = -0.051x + 41.076**	**Bias = 18 l/min**	**ICC_A = 0.891**
	r = 0.9307	PEFR = 223 to 687 l/min	LoA = -50 to 86	CI (0.795 to 0.935)
	y_2 = x	Error = 30 to 6 l/min	Lower limit CI	
	Slopes – equal	Mean PEFR= 453 l/min	= (-56 to -40)	
	Intercepts – not equal	Mean error = 18 l/min	Upper limit CI	
			= (74 to 90)	
	Agreement - NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**
10				
	y_1 = 0.963x + 33.768	**Error = -0.037x + 33.768**	**Bias = 17 l/min**	**ICC_A = 0.914**
	r = 0.9124	PEFR = 223 to 687 l/min	LoA = -48 to 82	CI (0.835 to 0.949)
	y_2 = x	Error = 26 to 8 l/min	Lower limit CI	
	Slopes – equal	Mean PEFR= 453 l/min	= (-59 to -42)	
	Intercepts – not equal	Mean error = 17 l/min	Upper limit CI	
			= (78 to 95)	
	Agreement - NO	**Agreement - YES**	**Agreement - NO**	**Agreement - YES**

Clinically significant difference for PEFR = 40l/min (Quanjer PH et al., 1997)

Table 4.22: Summary of prediction of agreement for all 10 clinical data set.

Method	Glucose level	SBP	DBP	Weight	PEFR
1. Comparison of slopes and y-intercepts analysis	1-YES 9-NO	7-YES 3-NO	10-NO	10-YES	10-NO
2. Agreement Model	10-YES	10-YES	10-YES	10-YES	10-YES
3. Bland-Altman LoA	10-NO	10-NO	10-NO	10-NO	10-NO
4. ICC_A	10-YES	10-YES	10-YES	10-YES	10-YES

NO - No agreement between instruments measuring the variable.
YES – Agreement between instruments measuring the variable.

4.3.2 Comparison of statistical methods for Agreement analysis: Simulated data

In this section, simulated data were set to represent a disagreement (with various range and distribution of error) of instrument measuring all variables (glucose level, SBP, DBP, weight, and PEFR). Any methods able to detect the bias and conclude disagreements in the analysis were considered to be correctly predict the disagreement. Section 4.3.2.1, Section 4.3.2.2 and Section 4.3.2.3 aim to determine how proportion and pattern of bias affect the prediction for each method. Whereas Section 4.3.2.4 aims to see the effect of sample size on the prediction of agreement for each method.

4.3.2.1 Constant systematic error

a. Overestimation of value (positive error)

All four methods correctly predict the disagreement (i.e. the presence of bias) in the data set of all variables. The agreement model and the Bland-Altman LoA predict the simulated positive error in all the data set. The y-intercept of the regression line also suggests the positive error in all the data set. The actual value of ICC_A did not provide good prediction of agreement, but the confidence interval (CI) reflects the disagreement in the dataset. The information from ICC_A analysis also does not provide information on the direction of error (positive or negative error) in the dataset. The summaries of analysis for all variables are displayed in Table 4.23.

Table 4.23: Comparison of agreement analysis with constant positive error

Variable	Comparison of slopes and y-intercepts analysis	Agreement Model	Bland-Altman LoA	ICC_A
Glucose x=Lab value, y_1= Lab value+ 0.8mmol/l	**$y_1 = x + 0.8$** **r(Pearson) = 1.0** **$y_2 = x$** Slopes – equal Intercepts – not equal **Agreement - NO**	**Error = 0.8** *Constant Error = 0.8mmol/l **Agreement - NO**	**Bias = 0.8mmol/l** **LoA = 0.8 to 0.8** *Constant Error = 0.8mmol/l **Agreement - NO**	**ICC_A = 0.911** CI (0.01 to 0.981) **Agreement - NO**
SBP x=SBP manual y_1= SBP manual + 10mmHg	**$y_1 = x + 10$** **r(Pearson) = 1.0** **$y_2 = x$** Slopes – equal Intercepts – not equal **Agreement - NO**	**Error = 10** *Constant Error = 10mmHg **Agreement - NO**	**Bias = 10mmHg** **LoA = 10 to 10** *Constant Error = 10mmHg **Agreement - NO**	**ICC_A = 0.865** CI (0.06 to 0.97) **Agreement - NO**
DBP x=DBP manual y_1= DBP manual + 10mmHg	**$y_1 = x + 10$** **r(Pearson) = 1.0** **$y_2 = x$** Slopes – equal Intercepts – not equal **Agreement - NO**	**Error = 10** *Constant Error = 10mmHg **Agreement - NO**	**Bias = 10mmHg** **LoA = 10 to 10** *Constant Error = 10mmHg **Agreement - NO**	**ICC_A = 0.749** CI (0.003 to 0.938) **Agreement - NO**
Weight x= Weight digital y_1= Weight digital + 0.5kg	**$y_1 = x + 0.5$** **r(Pearson) = 1.0** **$y_2 = x$** Slopes – not equal **Agreement - NO**	**Error = 0.5** *Constant Error =0.5kg **Agreement - NO**	**Bias = 0.5kg** **LoA = 0.5 to 0.5** *Constant Error =0.5kg **Agreement - NO**	**ICC_A = 0.999** CI (0.573 to 1.0) **Agreement - NO**
PEFR x=PEFR Clement Clarke UK average y_1= PEFR Clement Clarke UK average + 40 l/min	**$y_1 = x + 40$** **r(Pearson) = 1.0** **$y_2 = x$** Slopes – equal Intercept – not equal **Agreement - NO**	**Error = 40** *Constant Error = 40 l/min **Agreement - NO**	**Bias = 40 l/min** **LoA = 40 to 40** *Constant Error = 40 l/min **Agreement - NO**	**ICC_A = 0.892** CI (0.008 to 0.977) **Agreement - NO**

b. Underestimation of value (negative error)

All four methods correctly predict the disagreement in the data set of all variables. The agreement model and the Bland-Altman LoA predict the simulated negative error in all the data set. The y-intercept of the regression line also suggests the negative error in all the data set. The actual value of ICC_A did not provide good prediction of agreement, but the confidence interval (CI) reflects the disagreement in the dataset. The information from ICC_A analysis also does not provide information on the direction of error (positive or negative error) in the dataset. The summaries of analysis for all variables are displayed in Table 4.24.

Table 4.24: Comparison of agreement analysis with constant negative error

Variable	Comparison of slopes and y-intercepts analysis	Agreement Model	Bland-Altman LoA	ICC_A
Glucose x=Lab value, y_1= Lab value-0.8mmol/l	**$y_1 = x - 0.8$** **r(Pearson) = 1.0** **$y_2 = x$** Slopes – not equal **Agreement - NO**	**Error = -0.8** *Constant Error = -0.8mmol/l **Agreement – NO**	**Bias = -0.8mmol/l** **LoA = -0.8 to -0.8** *Constant Error = -0.8mmol/l **Agreement - NO**	**ICC = 0.911** CI (0.01 to 0.981) **Agreement - NO**
SBP x=SBP manual y_1= SBP manual - 10mmHg	**$y_1 = x - 10$** **r(Pearson) = 1.0** **$y_2 = x$** Slopes – equal Intercepts – not equal **Agreement - NO**	**Error = -10** *Constant Error = -10mmHg **Agreement - NO**	**Bias = -10mmHg** **LoA = -10 to -10** *Constant Error = -10mmHg **Agreement - NO**	**ICC_A = 0.865** CI (0.06 to 0.97) **Agreement - NO**
DBP x=DBP manual y_1= DBP manual -10mmHg	**$y_1 = x - 10$** **r(Pearson) = 1.0** **$y_2 = x$** Slopes – equal Intercepts – not equal **Agreement - NO**	**Error = -10** *Constant Error = -10mmHg **Agreement - NO**	**Bias = -10mmHg** **LoA = -10 to -10** *Constant Error = -10mmHg **Agreement - NO**	**ICC_A = 0.749** CI (0.003 to 0.938) **Agreement - NO**
Weight x= Weight digital y_1= Weight digital - 0.5kg	**$y_1 = x - 0.5$** **r(Pearson) = 1.0** **$y_2 = x$** Slopes – equal Intercept not equal **Agreement - NO**	**Error = -0.5** *Constant Error = -0.5kg **Agreement - NO**	**Bias = -0.5kg** **LoA = -0.5 to -0.5** *Constant Error = -0.5kg **Agreement - NO**	**ICC_A = 0.999** CI (0.573 to 1.0) **Agreement - NO**
PEFR x=PEFR Clement Clarke UK average y_1= PEFR Clement Clarke UK average - 40 l/min	**$y_1 = x - 40$** **r(Pearson) = 1.0** **$y_2 = x$** Slopes – equal Intercept – not equal **Agreement - NO**	**Error = 40** *Constant Error = -40 l/min **Agreement - NO**	**Bias = -40 l/min** **LoA = -40 to -40** *Constant Error = -40 l/min **Agreement - NO**	**ICC_A= 0.892** CI (0.008 to 0.977) **Agreement - NO**

4.3.2.2 Inconsistent error (mixture of positive and negative errors)

a. One-third positive error and two-thirds negative error

The comparison of slopes and y-intercepts analysis and the Bland-Altman LoA correctly predict the disagreement of data set for all variables. The agreement model and ICC_A detect the disagreement for some of the variables only. Although the Bland-Altman LoA successfully detects the disagreement in all data set, the actual biases predicted were overestimated. The overestimations of bias were seen for all variables, as shown in the Table 4.25. One example that demonstrates this is in the prediction of bias for the blood glucose level. The simulated bias for blood glucose level was ±0.8mmol/l, however the Bland-Altman LoA predicted that the bias was between -1.75mmol and 1.21mmol. In the simulated data set, one-third of the errors were positive and two-third of the errors were negative error. However, the mean bias predicted by the Bland-Altman method, agreement model and the y-intercept of the regression line show a negative error for all variables.

Table 4.25: Comparison of agreement analysis with 1/3 positive and 2/3 negative error

Variable	Comparison of slopes and y-intercepts analysis	Agreement Model	Bland-Altman LoA	ICC_A
Glucose x=Lab value, y_1= Lab value+ error Error = ±0.8mmol/l	**y_1 = 1.0006x -0.3** **r = 0.9240** **y_2 = x** Slopes – equal Intercepts – not equal	**Error = 0.0006x-0.3** Glucose=3.7to 14.2mmol/l Error=-0.30 to -0.29 mmol/l Mean glucose = 5.6 mmo/l Mean error = -0.30mmol/l	**Bias=-0.27 mmol/l** LoA= -1.75 to 1.21	**ICC_A = 0.912** CI (0.873 to 0.937)
	Agreement - NO	**Agreement – YES**	**Agreement - NO**	**Agreement– YES**
SBP x=SBP manual y_1= SBP manual + error Error = ±10mmHg	**y_1 = 1.359x – 47.374** **r = 0.9619** **y_2 = x** Slopes – not equal	**Error = 0.359x – 47.374** SBP = 84 to 192mmHg Error = -26 to 2mmHg Mean SBP = 123mmHg Mean error = -16 mmHg	**Bias = -3mmHg** **LoA = -22 to 15**	**ICC_A = 0.897** CI (0.853 to 0.926)
	Agreement - NO	**Agreement – NO**	**Agreement - NO**	**Agreement– YES**
DBP x=DBP manual y_1= DBP manual +error Error = ±10mmHg	**y_1 = 0.602x + 27.376** **r = 0.6723** **y_2 = x** Slopes – not equal	**Error = -0.398x+27.376** DBP = 44 to 110 mmHg Error = 10 to -16 mmHg Mean DBP = 77mmHg Mean error = -3 mmHg	**Bias = -3mmHg** **LoA = -22 to 15**	**ICC_A = 0.642** CI (0.533 to 0.724)
	Agreement – NO	**Agreement – NO**	**Agreement - NO**	**Agreement – NO**
Weight x= Weight digital y_1= Weight digital + error Error=± 0.5kg	**y_1 = 1.004x -0.436** **r = 0.9993** **y_2 = x** Slopes – equal Intercept – not equal	**Error = 0.004x-0.436** Weight = 38.5 to 111.4kg Error = -0.3 to 0.01kg Mean weight = 70kg Mean error = -0.16kg	**Bias = -0.17kg** **LoA = -1.09 to 0.76**	**ICC_A = 0.999** CI (0.999 to 0.999)
	Agreement - NO	**Agreement – YES**	**Agreement - NO**	**Agreement– YES**
PEFR x=PEFR Clement Clarke UK average y_1= PEFR Clement Clarke UK average + error Error = ±40 l/min	**y_1 = 1.06x – 40.37** **r = 0.9173** **y_2 = x** Slopes – not equal	**Error = 0.06x+40.37** PEFR = 223 to 687 l/min Error = 54 to 82 l/min Mean PEFR= 453 l/min Mean error = 68 l/min	**Bias = -13 l/min** **LoA = -87 to 61**	**ICC_A = 0.898** CI (0.853 to 0.926)
	Agreement - NO	**Agreement – NO**	**Agreement - NO**	**Agreement– YES**

b. Fifty percent positive error and fifty percent negative error

The Bland-Altman LoA correctly predicts the disagreement of data set for all variables. The comparison of slopes and y-intercepts analysis, agreement model and ICC_A only detect the disagreement for some of the variables only. Although the Bland-Altman LoA successfully detects the disagreement in all data set, the actual biases predicted were overestimated. One example is in the prediction of bias for the systolic blood pressure. The simulated bias for systolic blood pressure was ±10mmHg; however the Bland-Altman LoA predicted that the bias was between -20mmHg and 20mmHg. The overestimations of bias by the Bland-Altman LoA were seen for all variables (Table 4.26). In the simulated data set, the proportions of negative and positive error were equal. The mean bias predicted by the Bland-Altman method and mean error predicted by the agreement model for all variables were 0mmol/l (i.e. no error).

Table 4.26: Comparison of agreement analysis with 1/2 positive and 1/2 negative error

Variable	Comparison of slopes and y-intercepts analysis	Agreement Model	Bland-Altman LoA	ICC_A
Glucose x=Lab value, y_1= Lab value+ error Error = ±0.8mmol/l	**y_1 = 1.008x -0.044** **r = 0.9160** **y_2 = x** Slopes – equal Intercepts – equal **Agreement - YES**	**Error = 0.008x-0.044** Glucose=3.7to 14.2mmol/l Error=-0.01 to 0.07 mmol/l Mean glucose = 5.6 mmo/l Mean error = 0mmol/l **Agreement – YES**	**Bias= 0 mmol/l** LoA = -1.571 to 1.571 **Agreement - NO**	**ICC_A = 0.912** CI (0.873 to 0.937) **Agreement– YES**
SBP x=SBP manual y_1= SBP manual + error Error = ±10mmHg	**y_1 = 1.37x – 45.396** **r = 0.9562** **y_2 = x** Slopes – not equal **Agreement - NO**	**Error = 0.37x – 45.396** SBP = 84 to 192mmHg Error = -14 to 26mmHg Mean SBP = 123mmHg Mean error = 0 mmHg **Agreement – NO**	**Bias = 0mmHg** **LoA = -20 to 20** **Agreement - NO**	**ICC_A = 0.898** CI (0.873 to 0.918) **Agreement– YES**
DBP x=DBP manual y_1= DBP manual +error	**y_1 = 0.538x + 35.659** **r = 0.6218** **y_2 = x**	**Error = -0.462x+35.659** DBP = 44 to 110 mmHg Error = 15 to -15 mmHg	**Bias = 0mmHg** **LoA = -20 to 20**	**ICC_A = 0.616** CI (0.541 to 0.682)

Error = ±10mmHg	Slopes – not equal	Mean DBP = 77mmHg Mean error = 0 mmHg		
	Agreement – NO	**Agreement – NO**	**Agreement - NO**	**Agreement – NO**
Weight x= Weight digital y_1= Weight digital + error Error=± 0.5kg	**y_1 = 1.004x -0.285** **r = 0.9993** **y_2 = x** Slopes – equal Intercept – equal	**Error = 0.004x-0.285** Weight = 38.5 to 111.4kg Error = -0.13 to 0.16kg Mean weight = 70kg Mean error = 0kg	**Bias = 0kg** **LoA = -0.98 to 0.98**	**ICC_A = 0.999** CI (0.999 to 0.999)
	Agreement - YES	**Agreement – YES**	**Agreement - NO**	**Agreement– YES**
PEFR x=PEFR Clement Clarke UK average y_1= PEFR Clement Clarke UK average + error Error = ±40 l/min	**y_1 = 1.048x – 21.954** **r = 0.9061** **y_2 = x** Slopes – equal Intercept – equal	**Error = 0.048x – 21.954** PEFR = 223 to 687 l/min Error = -11 to 11 l/min Mean PEFR= 453 l/min Mean error = 0 l/min	**Bias = 0 l/min** **LoA = -79 to 79**	**ICC_A = 0.897** CI (0.872 to 0.917)
	Agreement - YES	**Agreement – YES**	**Agreement - NO**	**Agreement– YES**

c. Two-thirds positive error and one-third negative error

The comparison of slopes and y-intercepts analysis and the Bland-Altman LoA correcltly predict the disagreement of data set for all variables. The agreement model and ICC_A detect the disagreement for some of the variables only. The actual biases predicted by Bland-Altman method were overestimated. The overestimations of bias were seen in all variables (see Table 4.27). One example is the result for diastolic pressure. The simulated bias for diastolic blood pressure was ±10mmHg, however the Bland-Altman LoA predicted that the bias was between -15mmHg and 22mmHg. In the simulated data set, two-third of the errors were positive and one-third of the errors were negative error. The mean bias predicted by the Bland-Altman method and agreement model show a positive error for all variables.

Table 4.27: Comparison of agreement analysis with 2/3 positive and 1/3 negative error

Variable	Comparison of slopes and y-intercepts analysis	Agreement Model	Bland-Altman LoA	ICC_A
Glucose x=Lab value, y_1= Lab value+ error Error = ±0.8mmol/l	**y_1 = 1.018x -0.3** **r = 0.9257** **y_2 = x** Slopes – equal Intercepts – not equal **Agreement - NO**	**Error = 0.018x+0.167** Glucose=3.7to 14.2mmol/l Error= 0.23 to 0.42 mmol/l Mean glucose = 5.6 mmo/l Mean error = 0.27mmol/l **Agreement – YES**	**Bias= 0.27 mmol/l** LoA= -1.21 to 1.75 **Agreement - NO**	**ICC_A = 0.913** CI (0.874 to 0.938) **Agreement– YES**
SBP x=SBP manual y_1= SBP manual + error Error = ±10mmHg	**y_1 = 1.306x – 34.211** **r = 0.9500** **y_2 = x** Slopes – not equal **Agreement - NO**	**Error = 0.306x – 34.211** SBP = 84 to 192mmHg Error = -9 to 25mmHg Mean SBP = 123mmHg Mean error = 3 mmHg **Agreement – NO**	**Bias = 3mmHg** **LoA = -15 to 22** **Agreement - NO**	**ICC_A = 0.893** CI (0.847 to 0.923) **Agreement– YES**
DBP x=DBP manual y_1= DBP manual +error Error = ±10mmHg	**y_1 = 0.561x + 37.376** **r = 0.6613** **y_2 = x** Slopes – not equal **Agreement – NO**	**Error = -0.439x+37.376** DBP = 44 to 110 mmHg Error = 18 to -11 mmHg Mean DBP = 77mmHg Mean error = 4 mmHg **Agreement – NO**	**Bias = 3mmHg** **LoA = -15 to 22** **Agreement - NO**	**ICC_A = 0.626** CI (0.515 to 0.711) **Agreement – NO**
Weight x= Weight digital y_1= Weight digital + error Error=± 0.5kg	**y_1 = 1.003x -0.03** **r = 0.9993** **y_2 = x** Slopes – equal Intercept – not equal **Agreement - NO**	**Error = 0.003x-0.03** Weight = 38.5 to 111.4kg Error = 0.09 to 0.30kg Mean weight = 70kg Mean error = 0.18kg **Agreement – YES**	**Bias = 0.17kg** **LoA=-0.76 to 1.09** **Agreement - NO**	**ICC_A = 0.999** CI (0.999 to 0.999) **Agreement– YES**
PEFR x=PEFR Clement Clarke UK average y_1= PEFR Clement Clarke UK average + error Error = ±40 l/min	**y_1 = 1.069x – 18.053** **r = 0.9190** **y_2 = x** Slopes – not equal **Agreement - NO**	**Error = 0.069x– 18.053** PEFR = 223 to 687 l/min Error = -3 to 29 l/min Mean PEFR= 453 l/min Mean error = 13 l/min **Agreement – YES**	**Bias = 13 l/min** **LoA = -61 to 87** **Agreement - NO**	**ICC_A = 0.899** CI (0.855 to 0.927) **Agreement– YES**

d. One-third positive error, one-third negative error and one-third agreement.

The Bland-Altman LoA correctly identifies the disagreement of data set for all variables. The comparison of slopes and intercepts analysis, agreement model and ICC_A detect the disagreement for some of the variables only. The Bland-Altman LoA overestimates the actual biases predicted. As an example, the simulated bias for body weight was ±0.5kg, however the Bland-Altman LoA predicted that the bias was between -0.8kg and 0.8kg (Table 4.28). The overestimations of bias by the Bland-Altman LoA were seen in all variables, and the results are summarised in Table 4.28. In the simulated data set, the proportions of agreement, negative error and positive error were equal. The mean bias predicted by the Bland-Altman method for all variables were 0mmol/l (i.e. no error).

Table 4.28: Comparison of agreement analysis with 1/3 positive error, 1/3 negative error, and 1/3 agreement.

Variable	Comparison of slopes and y-intercepts analysis	Agreement Model	Bland-Altman LoA	ICC_A
Glucose x=Lab value, y_1= Lab value+ error Error = ±0.8mmol/l	**y_1 = 0.997x -0.016** **r = 0.9404** **y_2 = x** Slopes – equal Intercepts – equal **Agreement – YES**	**Error = -0.003x-0.016** Glucose=3.7to 14.2mmol/l Error=-0.03 to -0.06 mmol/l Mean glucose = 5.6 mmo/l Mean error = -0.03mmol/l **Agreement - YES**	**Bias= 0 mmol/l** LoA= -1.28 to 1.28 **Agreement - NO**	**ICC_A = 0.939** CI (0.924 to 0.951) **Agreement– YES**
SBP x=SBP manual y_1= SBP manual + error Error = ±10mmHg	**y_1 = 1.206x – 25.268** **r = 0.94736** **y_2 = x** Slopes – not equal **Agreement – NO**	**Error = 0.206x – 25.268** SBP = 84 to 192mmHg Error = -8 to 14mmHg Mean SBP = 123mmHg Mean error = 0 mmHg **Agreement - NO**	**Bias = 0mmHg** **LoA = -16 to 16** **Agreement - NO**	**ICC_A = 0.921** CI (0.901 to 0.936) **Agreement– YES**
DBP x=DBP manual y_1= DBP manual +error Error =	**y_1 = 0.578x + 32.618** **r = 0.7395** **y_2 = x** Slopes – not equal	**Error = -0.422x+32.618** DBP = 44 to 110 mmHg Error = 14 to -14 mmHg Mean DBP = 77mmHg Mean error = 0 mmHg	**Bias = 0mmHg** **LoA = -20 to 20**	**ICC_A = 0.616** CI (0.541 to 0.682)

±10mmHg	**Agreement – NO**	**Agreement – NO**	**Agreement - NO**	**Agreement – NO**
Weight x= Weight digital y_1= Weight digital + error Error=± 0.5kg	**y_1 = 1.003x -0.233** **r = 0.9995** **y_2 = x** Slopes – equal Intercept – equal **Agreement – YES**	**Error = 0.003x-0.233** Weight = 38.5 to 111.4kg Error = -0.12 to 0.10kg Mean weight = 70kg Mean error = -0.02kg **Agreement – YES**	**Bias = 0kg** **LoA = -0.80 to 0.80** **Agreement - NO**	**ICC_A = 1.0** CI (0.999 to 0.999) **Agreement– YES**
PEFR x=PEFR Clement Clarke UK average y_1= PEFR Clement Clarke UK average + error Error = ±40 l/min	**y_1 = 1.064x – 29.212** **r = 0.9371** **y_2 = x** Slopes – not equal	**Error = 0.064x – 29.212** PEFR = 223 to 687 l/min Error = -15 to 15 l/min Mean PEFR= 453 l/min Mean error = 0 l/min	**Bias = 0 l/min** **LoA = -64 to 64**	**ICC_A = 0.930** CI (0.913 to 0.944)
	Agreement – NO	**Agreement - YES**	**Agreement - NO**	**Agreement– YES**

4.3.2.3 Proportion of error

a. One third of positive error, and two third of agreement in dataset

The comparison of slopes and y-intercepts analysis and the Bland-Altman LoA correctly predict the disagreement in the data set for all variables. The comparison of slopes and y-intercepts analysis did not provide information on the direction or quantification of error for all the variables. The mean error (predicted in the agreement model) and the mean bias (predicted by the Bland-Altman method) were lower than the actual error in the simulated data set for all the variables. The agreement model only detects the disagreement in the data set for systolic blood pressure and diastolic blood pressure. The ICC_A only detect disagreement in the data set for diastolic blood pressure. The summary of agreement analysis of data set with one-third of error for all variables is display in the Table 4.29.

Table 4.29: Comparison of agreement analysis with 1/3 error

Variable	Comparison of slopes and y-intercepts analysis	Agreement Model	Bland-Altman LoA	ICC_A
Glucose x=Lab value, y_1= Lab value+ error Error = +0.8mmol/l	**y_1 = 0.991x +0.317** **r = 0.9787** **y_2 = x** Slopes – equal Intercepts – not equal **Agreement – NO**	**Error = -0.009x +0.317** Glucose=3.7to 14.2mmol/l Error=0.28 to 0.19 mmol/l Mean glucose = 5.6 mmo/l Mean error = 0.27mmol/l **Agreement – YES**	**Bias=0.27 mmol/l** LoA= -0.47 to 1.01 **Agreement - NO**	**ICC_A = 0.968** CI (0.899 to 0.985) **Agreement– YES**
SBP x=SBP manual y_1= SBP manual + error Error = +10mmHg	**y_1 = 1.180x – 18.687** **r = 0.9869** **y_2 = x** Slopes – not equal **Agreement – NO**	**Error = 0.18x – 18.687** SBP = 84 to 192mmHg Error = -4 to 16mmHg Mean SBP = 123mmHg Mean error = 3 mmHg **Agreement - NO**	**Bias = 3mmHg** LoA = -6 to 13 **Agreement - NO**	**ICC_A = 0.958** CI (0.869 to 0.980) **Agreement– YES**
DBP x=DBP manual y_1= DBP manual +error Error = +10mmHg	**y_1 = 0.801x + 18.688** **r = 0.9241** **y_2 = x** Slopes – not equal **Agreement – NO**	**Error = -0.199x+18.688** DBP = 44 to 110 mmHg Error = 10 to -3 mmHg Mean DBP = 77mmHg Mean error = 3 mmHg **Agreement – NO**	**Bias = 3mmHg** LoA = -6 to 13 **Agreement - NO**	**ICC_A = 0.878** CI (0.670 to 0.939) **Agreement – NO**
Weight x= Weight digital y_1= Weight digital + error Error= +0.5kg	**y_1 = 1.002x -0.032** **r = 0.9998** **y_2 = x** Slopes – equal Intercept – not equal **Agreement – NO**	**Error = 0.002x-0.032** Weight = 38.5 to 111.4kg Error = 0.05 to 0.19kg Mean weight = 70kg Mean error = 0.11kg **Agreement – YES**	**Bias = 0.17kg** LoA= -0.30 to 0.63 **Agreement - NO**	**ICC_A = 1.0** CI (0.999 to 1.0) **Agreement– YES**
PEFR x=PEFR Clement Clarke UK average y_1= PEFR Clement Clarke UK average + error Error = +40 l/min	**y_1 = 1.03x – 0.185** **r = 0.9760** **y_2 = x** Slopes – not equal **Agreement – NO**	**Error = 0.030x – 0.185** PEFR = 223 to 687 l/min Error = 7 to 20 l/min Mean PEFR= 453 l/min Mean error = 13 l/min **Agreement - YES**	**Bias = 13 l/min** LoA = -24 to 50 **Agreement - NO**	**ICC_A = 0.962** CI (0.882 to 0.982) **Agreement– YES**

a. Fifty percent positive error, and fifty percent agreement in dataset

The comparing two straight lines analysis and the Bland-Altman LoA predict the disagreement in the data set for all variables. The comparing two straight lines analysis did not provide information on the direction or quantification of error for all the variables. The mean error (predicted in the agreement model) and the mean bias (predicted by the Bland-Altman method) were lower than the actual error in the simulated data set for all the variables. The ICC_A detects disagreement in the data set for systolic blood pressure, diastolic blood pressure and peak expiratory flow meter. The agreement model only detects the disagreement in the data set for systolic blood pressure and diastolic blood pressure. The summary of agreement analysis of data set with fifty percent of error for all variables is displayed in Table 4.30.

Table 4.30: Comparison of agreement analysis with 1/2 error

Variable	Comparison of slopes and y-intercepts analysis	Agreement Model	Bland-Altman LoA	ICC_A
Glucose x=Lab value, y_1= Lab value+ error Error = +0.8mmol/l	**y_1 = 0.996x +0.422** **r = 0.9763** **y_2 = x** Slopes – equal Intercepts – not equal	**Error = -0.004x+0.422** Glucose=3.7to 14.2mmol/l Error=0.41 to 0.37 mmol/l Mean glucose = 5.6 mmo/l Mean error = 0.40mmol/l	**Bias= 0.4 mmol/l** **LoA= -0.39 to 1.19**	**ICC_A = 0.954** CI (0.701 to 0.983)
	Agreement - NO	**Agreement – YES**	**Agreement - NO**	**Agreement– YES**
SBP x=SBP manual y_1= SBP manual + error Error = +10mmHg	**y_1 = 1.185x – 17.698** **r = 0.9847** **y_2 = x** Slopes – not equal	**Error = 0.185x – 17.698** SBP = 84 to 192mmHg Error = -2 to 18mmHg Mean SBP = 123mmHg Mean error = 5 mmHg	**Bias = 5mmHg** **LoA = -5 to 15**	**ICC_A = 0.938** CI (0.630 to 0.977)
	Agreement - NO	**Agreement – NO**	**Agreement - NO**	**Agreement– NO**
DBP x=DBP manual y_1= DBP manual +error Error = +10mmHg	**y_1 = 0.769x + 22.83** **r = 0.9152** **y_2 = x** Slopes – not equal	**Error = -0.231x+22.83** DBP = 44 to 110 mmHg Error = 13 to -3 mmHg Mean DBP = 77mmHg Mean error = 5 mmHg	**Bias = 5mmHg** **LoA = -5 to 15**	**ICC_A = 0.821** CI (0.301 to 0.928)
	Agreement – NO	**Agreement – NO**	**Agreement - NO**	**Agreement – NO**
Weight x= Weight digital y_1= Weight digital + error Error= +0.5kg	**y_1 = 1.002x +0.107** **r = 0.9998** **y_2 = x** Slopes – equal Intercept – not equal	**Error = 0.002x+0.107** Weight = 38.5 to 111.4kg Error = 0.18 to 0.33kg Mean weight = 70kg Mean error = 0.25kg	**Bias = 0.25kg** **LoA=-0.24 to 0.74**	**ICC_A = 1.0** CI (0.997 to 1.0)
	Agreement - NO	**Agreement – YES**	**Agreement - NO**	**Agreement– YES**
PEFR x=PEFR Clement Clarke UK average y_1= PEFR Clement Clarke UK average + error Error = +40 l/min	**y_1 = 1.024x + 9.023** **r = 0.9726** **y_2 = x** Slopes – equal Intercept – not equal	**Error = 0.024x + 9.023** PEFR = 223 to 687 l/min Error = 14 to 26 l/min Mean PEFR= 453 l/min Mean error = 20 l/min	**Bias = 20 l/min** **LoA = -20 to 60**	**ICC_A = 0.944** CI (0.657 to 0.9980)
	Agreement - NO	**Agreement – YES**	**Agreement - NO**	**Agreement– NO**

a. Two-thirds positive error, and one third agreement in dataset

The comparison of slopes and y-intercepts analysis and the Bland-Altman LoA show the disagreement in the data set for all variables. The comparison of slopes and y-intercepts analysis did not provide information on the direction or quantification of error for all the variables. The mean error (predicted in the agreement model) and the mean bias (predicted by the Bland-Altman method) were lower than the actual error in the simulated data set for all the variables. The ICC_A detects disagreement in the data set for all variables except for weight. The agreement model only detects the disagreement in the data set for systolic blood pressure and diastolic blood pressure. The summary of agreement analysis of data set with two-thirds error for all variables is displayed in Table 4.31.

Table 4.31: Comparison of agreement analysis with 2/3 error

Variable	Comparison of slopes and y-intercepts analysis	Agreement Model	Bland-Altman LoA	ICC_A
Glucose x=Lab value, y_1= Lab value+ error Error = +0.8mmol/l	**y_1 = 0.997x +0.55** **r = 0.9789** **y_2 = x** Slopes – equal Intercepts – not equal **Agreement - NO**	**Error = -0.003x+0.55** Glucose=3.7to 14.2mmol/l Error= 0.54 to 0.51 mmol/l Mean glucose = 5.6 mmo/l Mean error = 0.53mmol/l **Agreement - YES**	**Bias= 0.53 mmol/l** LoA= -0.21 to 1.27 **Agreement - NO**	**ICC_A = 0.939** CI (0.298 to 0.982) **Agreement– NO**
SBP x=SBP manual y_1= SBP manual + error Error = +10mmHg	**y_1 = 1.153x – 12.105** **r = 0.9831** **y_2 = x** Slopes – not equal **Agreement - NO**	**Error = 0.153x – 12.105** SBP = 84 to 192mmHg Error = 1 to 17mmHg Mean SBP = 123mmHg Mean error = 7 mmHg **Agreement - NO**	**Bias = 7 mmHg** **LoA = -3 to 16** **Agreement - NO**	**ICC_A = 0.917** CI (0.213 to 0.975) **Agreement– NO**
DBP x=DBP manual y_1= DBP manual +error Error = +10mmHg	**y_1 = 0.780x + 23.60** **r = 0.9260** **y_2 = x** Slopes – not equal **Agreement – NO**	**Error = -0.220x+23.60** DBP = 44 to 110 mmHg Error = 14 to -1 mmHg Mean DBP = 77mmHg Mean error = 7 mmHg **Agreement – NO**	**Bias = 7mmHg** **LoA = -3 to 16** **Agreement - NO**	**ICC_A = 0.777** CI (0.001 to 0.925) **Agreement – NO**
Weight x= Weight digital y_1= Weight digital + error Error= +0.5kg	**y_1 = 1.001x +0.235** **r = 0.9998** **y_2 = x** Slopes – equal Intercept – not equal **Agreement - NO**	**Error = 0.001x+0.235** Weight = 38.5 to 111.4kg Error = 0.27 to 0.35kg Mean weight = 70kg Mean error = 0.31kg **Agreement – YES**	**Bias = 0.33kg** **LoA = -0.13 to 0.80** **Agreement - NO**	**ICC_A = 1.0** CI (0.987 to 1.0) **Agreement– YES**
PEFR x=PEFR Clement Clarke UK average y_1= PEFR Clement Clarke UK average + error Error = +40 l/min	**y_1 = 1.035x +10.973** **r = 0.9763** **y_2 = x** Slopes – not equal **Agreement - NO**	**Error = 0.035x +10.973** PEFR = 223 to 687 l/min Error = 19 to 35 l/min Mean PEFR= 453 l/min Mean error = 27 l/min **Agreement - YES**	**Bias = 27 l/min** **LoA = -10 to 64** **Agreement - NO**	**ICC_A = 0.928** CI (0.250 to 0.979) **Agreement– NO**

4.3.2.4 Sample Size

The results of the analysis on the effect of sample size are described in this section. The analysis of each statistical method was run 10 times with sample size from 10 to 500 (random sampling with replacement) for all variables. The standard deviation and standard error of the prediction from the 10 sets of analysis were calculated. Results for the analysis of blood glucose level, systolic BP, diastolic BP, weight, and PEFR are shown in Figure 4.2 to Figure 4.6. The standard error of all outcomes for all tested methods decreases as the sample size increases, and stabilises after a certain sample size. The standard errors of the prediction become stable when the sample size is greater than 100. The pattern of prediction becomes more consistent after the sample size is greater than 200.

a. Blood glucose level

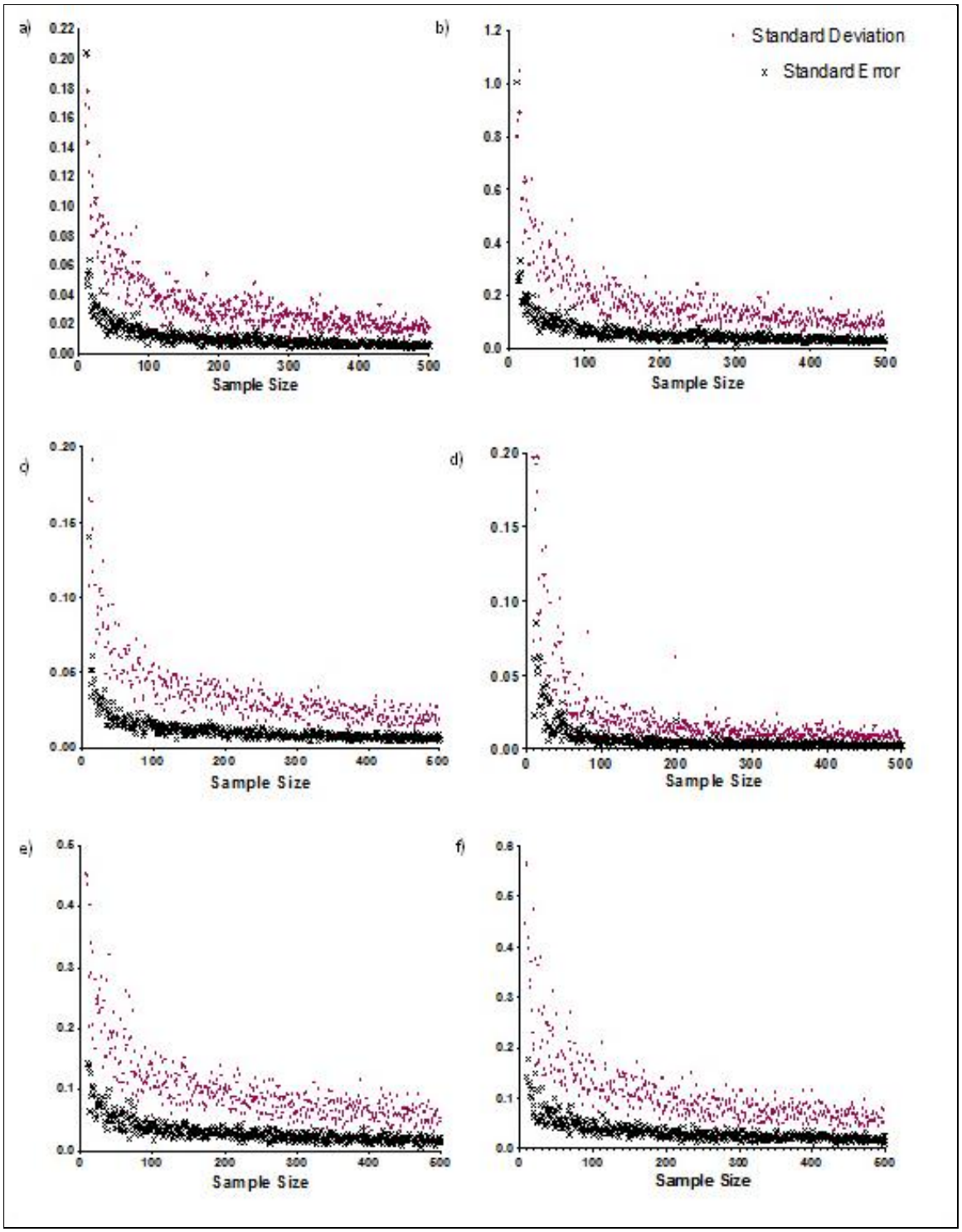


Figure 4.2: The effect of sample size on the prediction of (a) slope, (b) intercept, (c) agreement model, (d) ICC_A, (e) upper limit of agreement, and (f) lower limit of agreement for blood glucose level.

b. Systolic blood pressure (SBP)

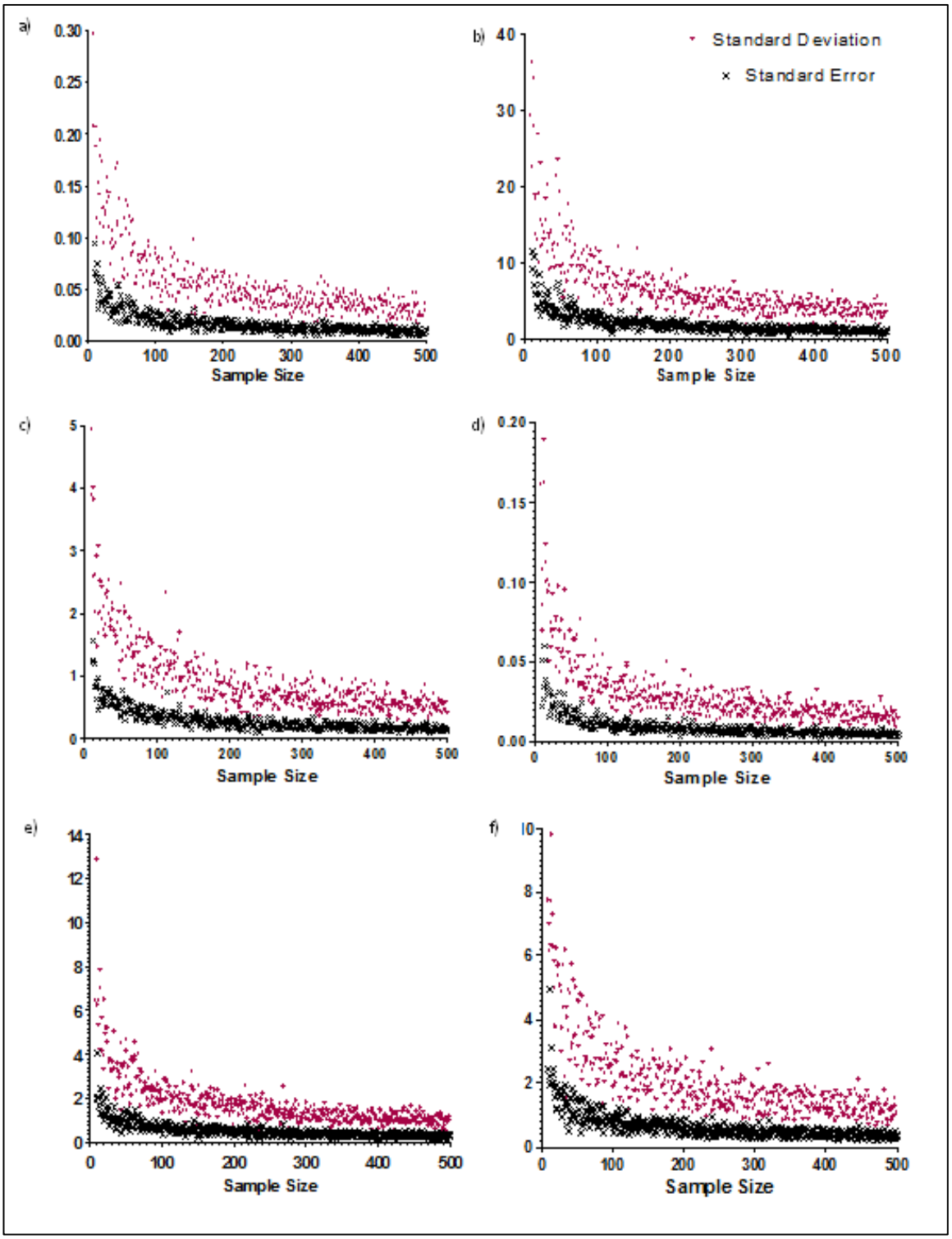


Figure 4.3: The effect of sample size on the prediction of (a) slope, (b) intercept, (c) agreement model, (d) ICC_A, (e) upper limit of agreement, and (f) lower limit of agreement for systolic blood pressure.

c. Diastolic blood pressure (DBP)

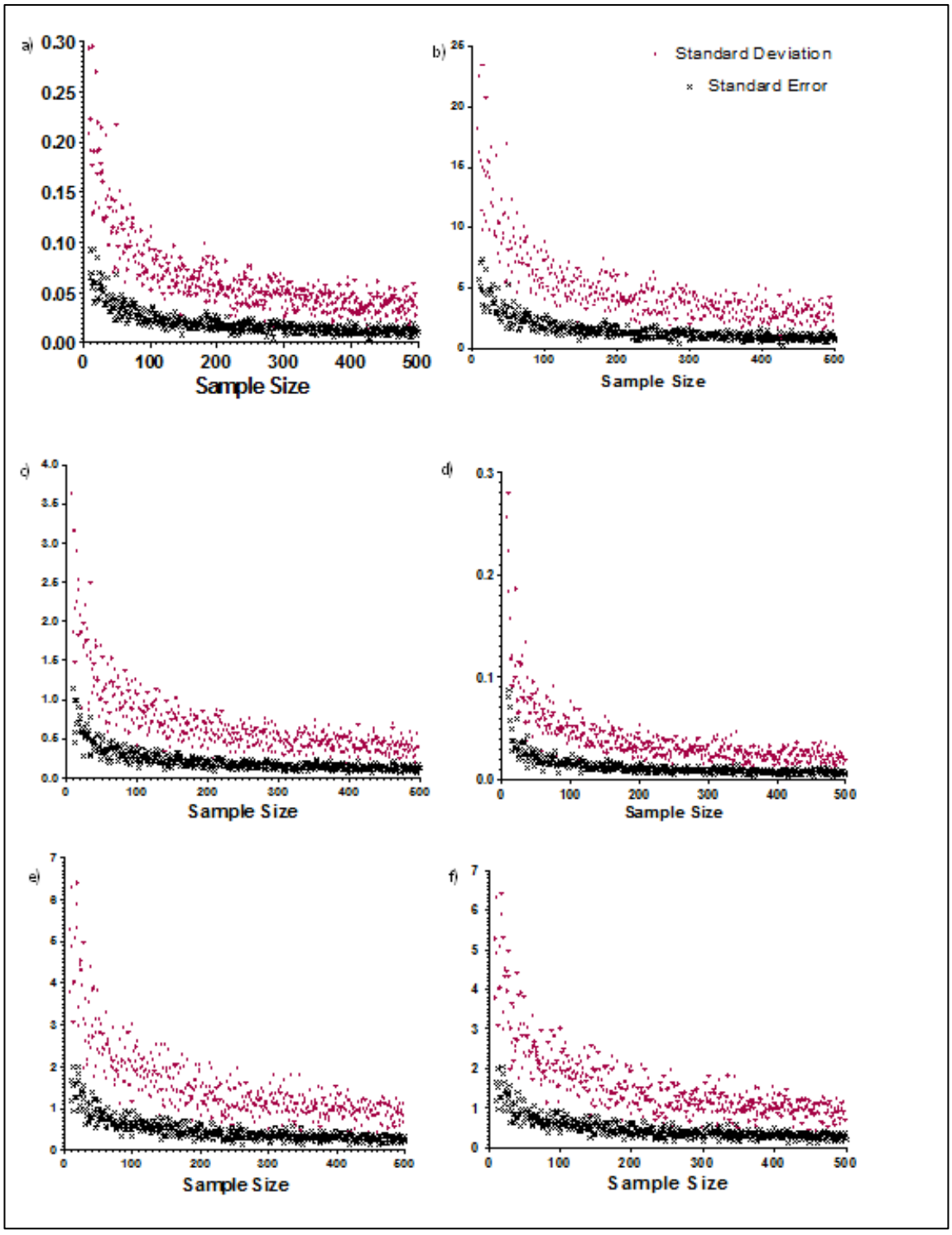


Figure 4.4: The effect of sample size on the prediction of (a) slope, (b) intercept, (c) agreement model, (d) ICC_A, (e) upper limit of agreement, and (f) lower limit of agreement for diastolic blood pressure.

d. Weight

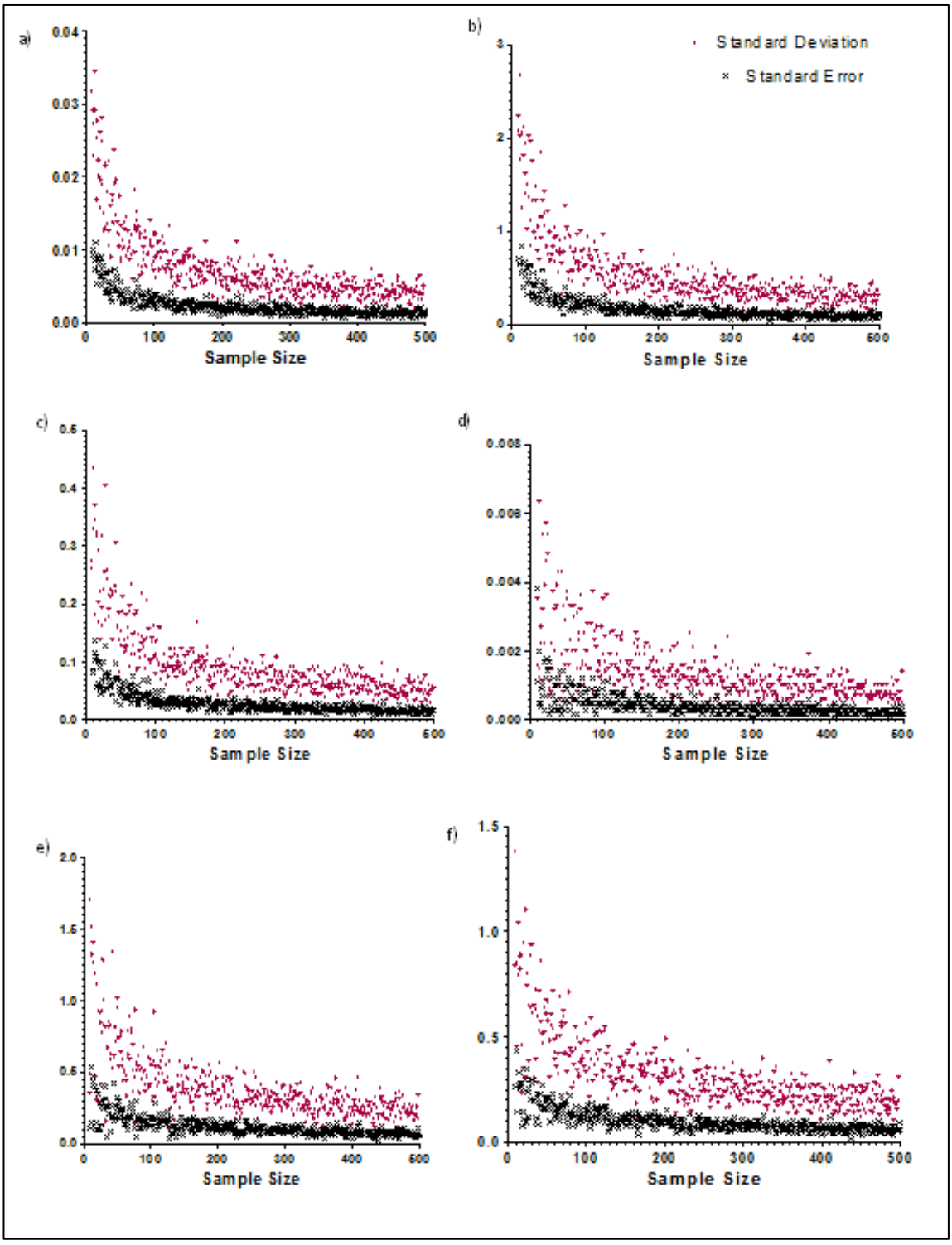


Figure 4.5: The effect of sample size on the prediction of (a) slope, (b) intercept, (c) agreement model, (d) ICC_A, (e) upper limit of agreement, and (f) lower limit of agreement for body weight.

e. Peak expiratory flow rate (PEFR)

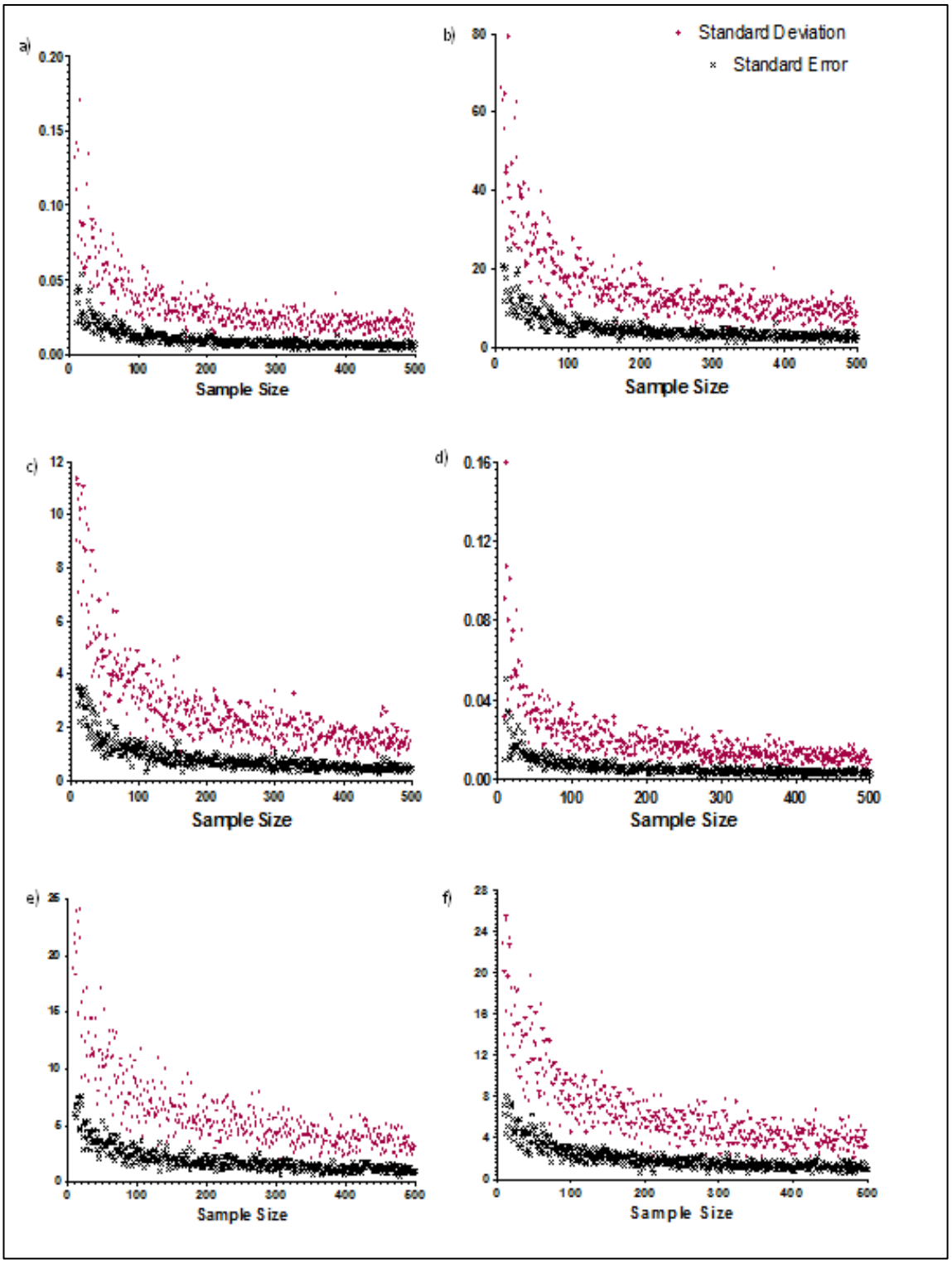


Figure 4.6: The effect of sample size on the prediction of (a) slope, (b) intercept, (c) agreement model, (d) ICC_A, (e) upper limit of agreement, and (f) lower limit of agreement for peak expiratory flow rate.

The impact of sample size on the prediction of bias can be demonstrated using the plot of bias versus the sample size. Figure 4.7 shows the result for the estimation of bias using agreement model for blood glucose level variable. The predictions are unstable when the sample size is less than 100. For instance, when the sample size is 25, the analysis of simulated data for blood glucose level shows that the prediction of error can be as low as 0.26mmol/l and as high as 0.41mmol/l. When the sample size is 75, the prediction of error is between 0.24mmol/l and 0.37mmol/l. The pattern is similar to the analysis for systolic BP (Figure 4.8). Similar trends were seen for the other variables (diastolic BP, weight and PEFR). Results for these variables are shown in the Figure 4.9.

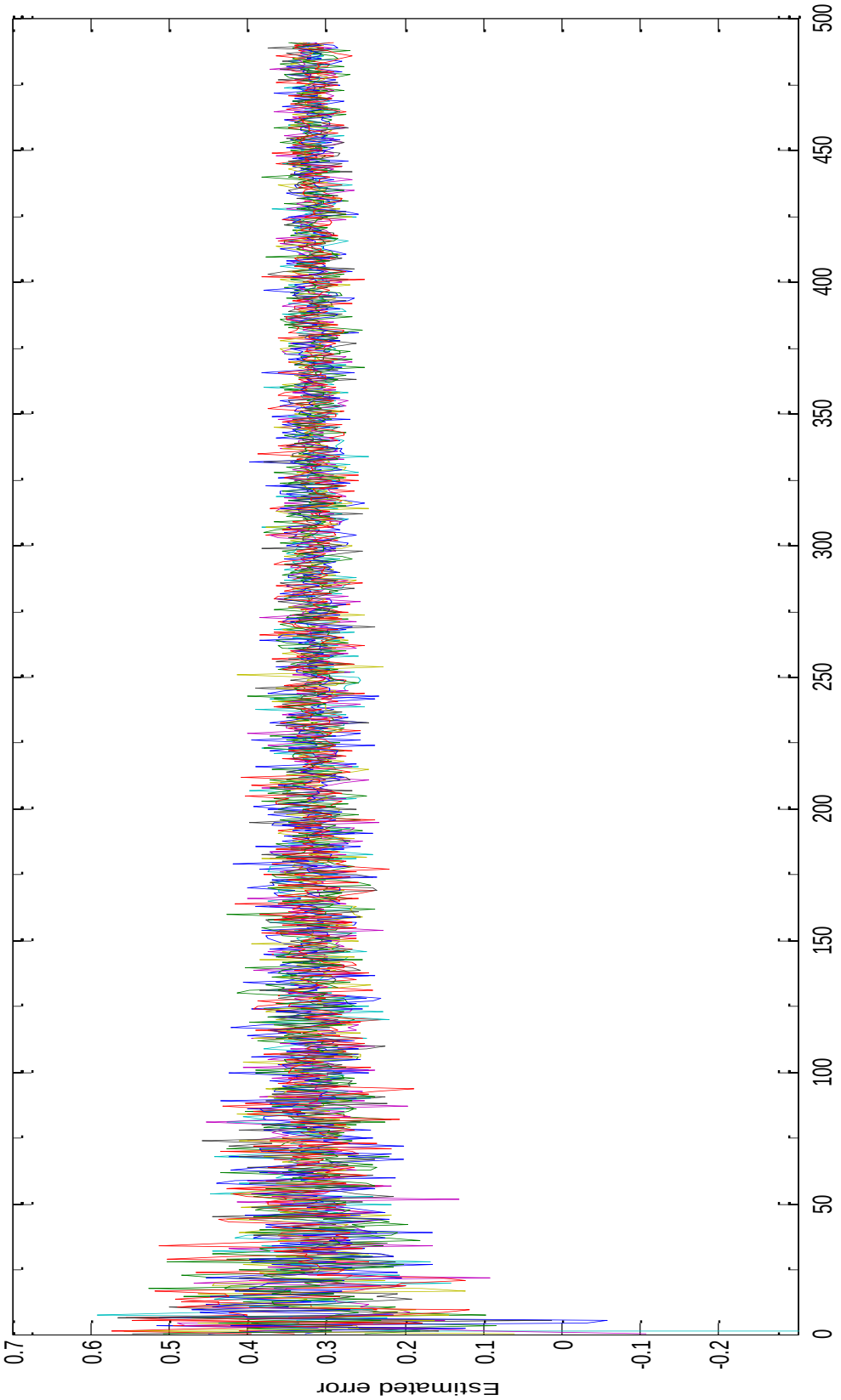


Figure 4.7: The plot of the prediction of error versus sample size for all 10 sets of data for glucose

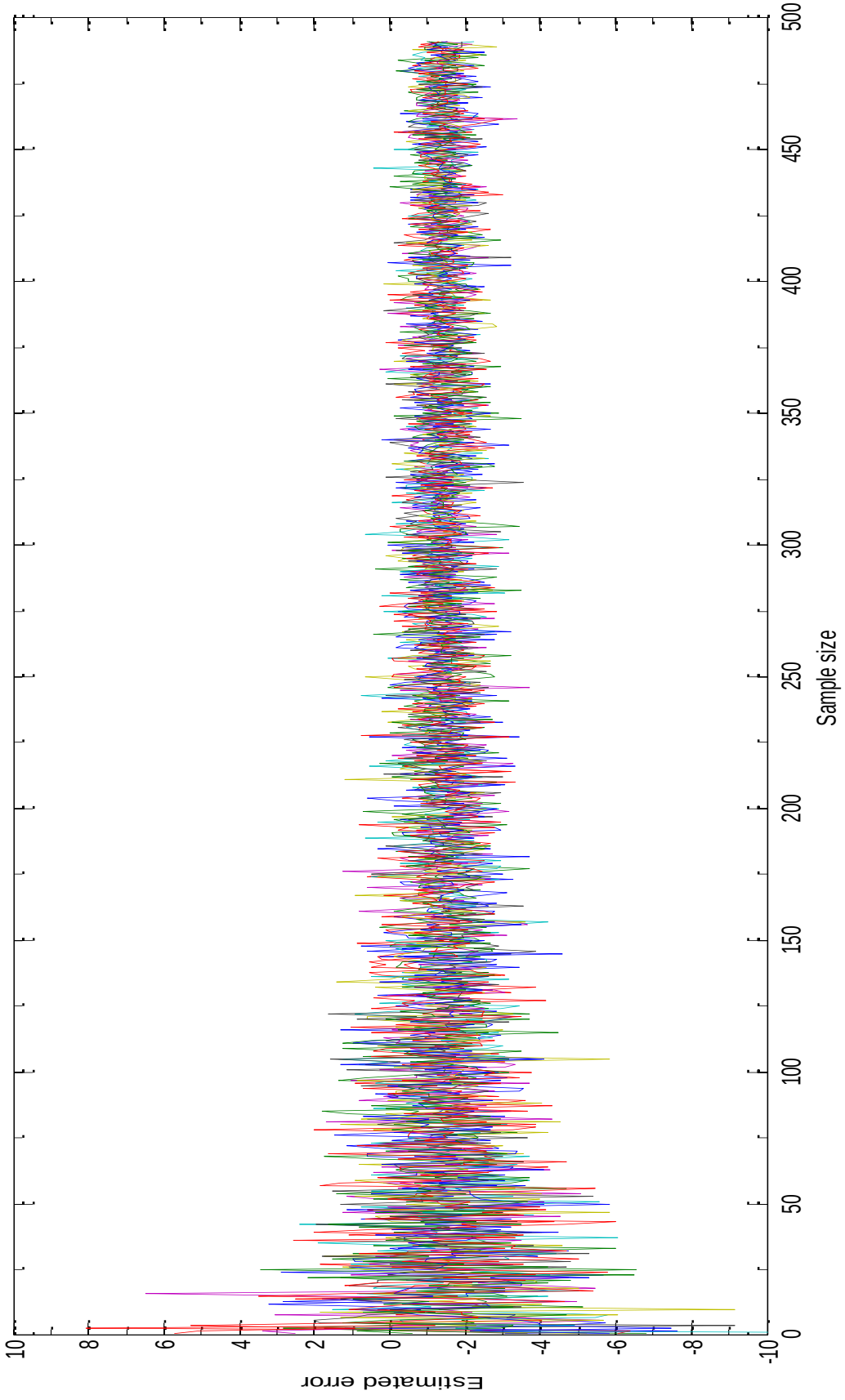


Figure 4.8: The plot of the prediction of error versus sample size for all 10 sets of data for SBP

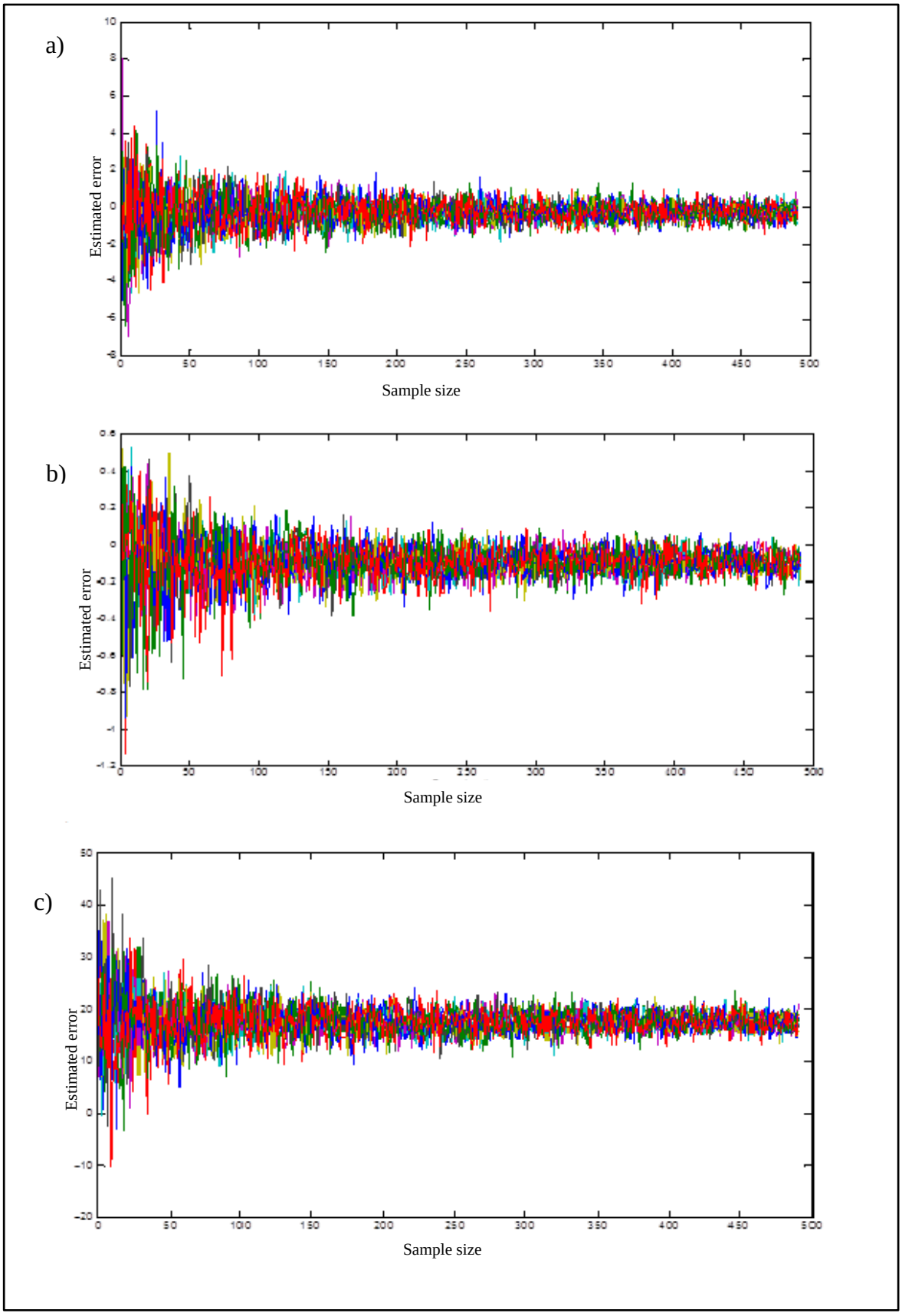


Figure 4.9: The plot of bias predicted using agreement model versus sample size for all 10 sets of analysis for (a) DBP, (b) weight, and (c) PEFR.

Figure 4.10 and Figure 4.11 shows the effect of sample size on the prediction of the Bland-Altman analysis for blood glucose level and diastolic BP. The prediction of bias and limits of agreement seems to be unstable especially when the sample size is less than 100. For instance, when the sample size is 30, the analysis of simulated data for blood glucose level shows that the prediction of lower limit of agreement can be as low as -1.59mmol/l and as high as -0.96mmol/l. The prediction of upper limit of agreement can be as low as 0.31mmol/l and the maximum prediction for upper limit of agreement is 1.03mmol/l. The pattern is similar to the analysis for diastolic BP, when the sample size is 30 simulated data shows that the prediction of lower limit of agreement can be between -23mmHg and -9mmHg. The prediction of upper limit of agreement can be between 11mmHg and 28mmHg. Similar trend were seen for the other variables (systolic BP, weight and PEFR). Results for these variables are shown in Figure 4.12.

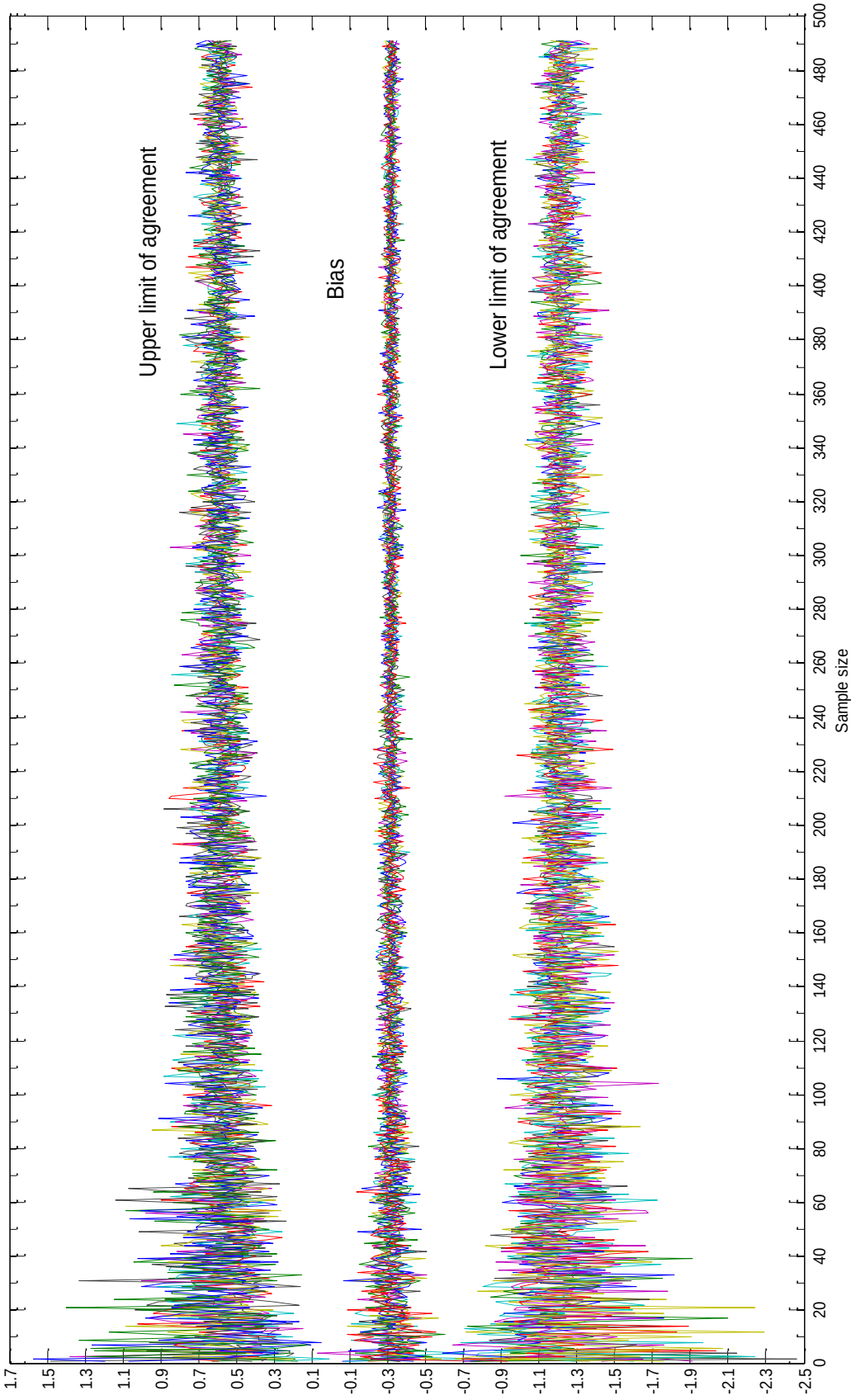


Figure 4.10: The plot of the prediction of bias and limits of agreement versus sample size for all 10 sets of data for glucose

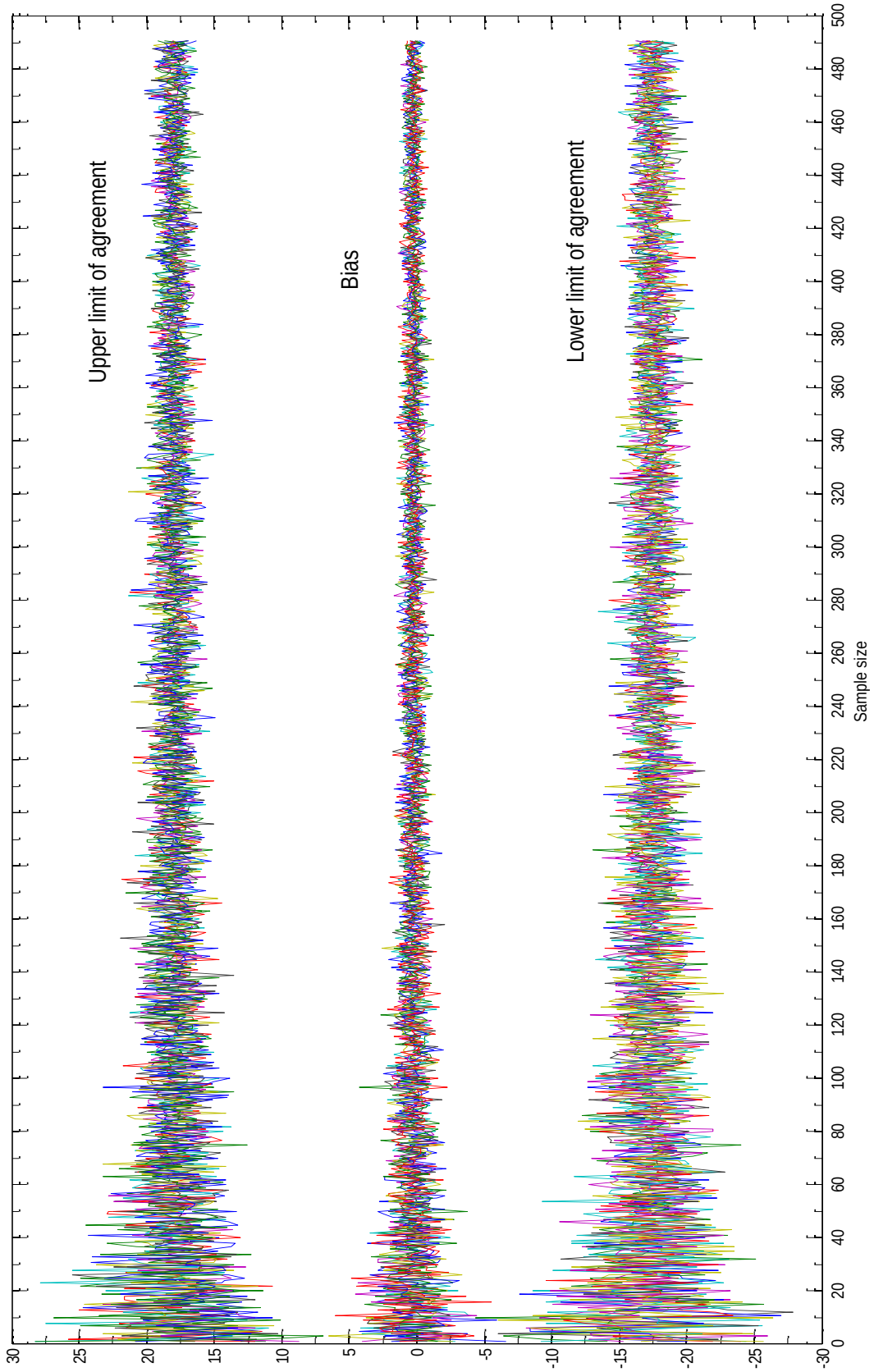


Figure 4.11: The plot of the prediction of bias and limits of agreement versus sample size for all 10 sets of data for diastolic BP.

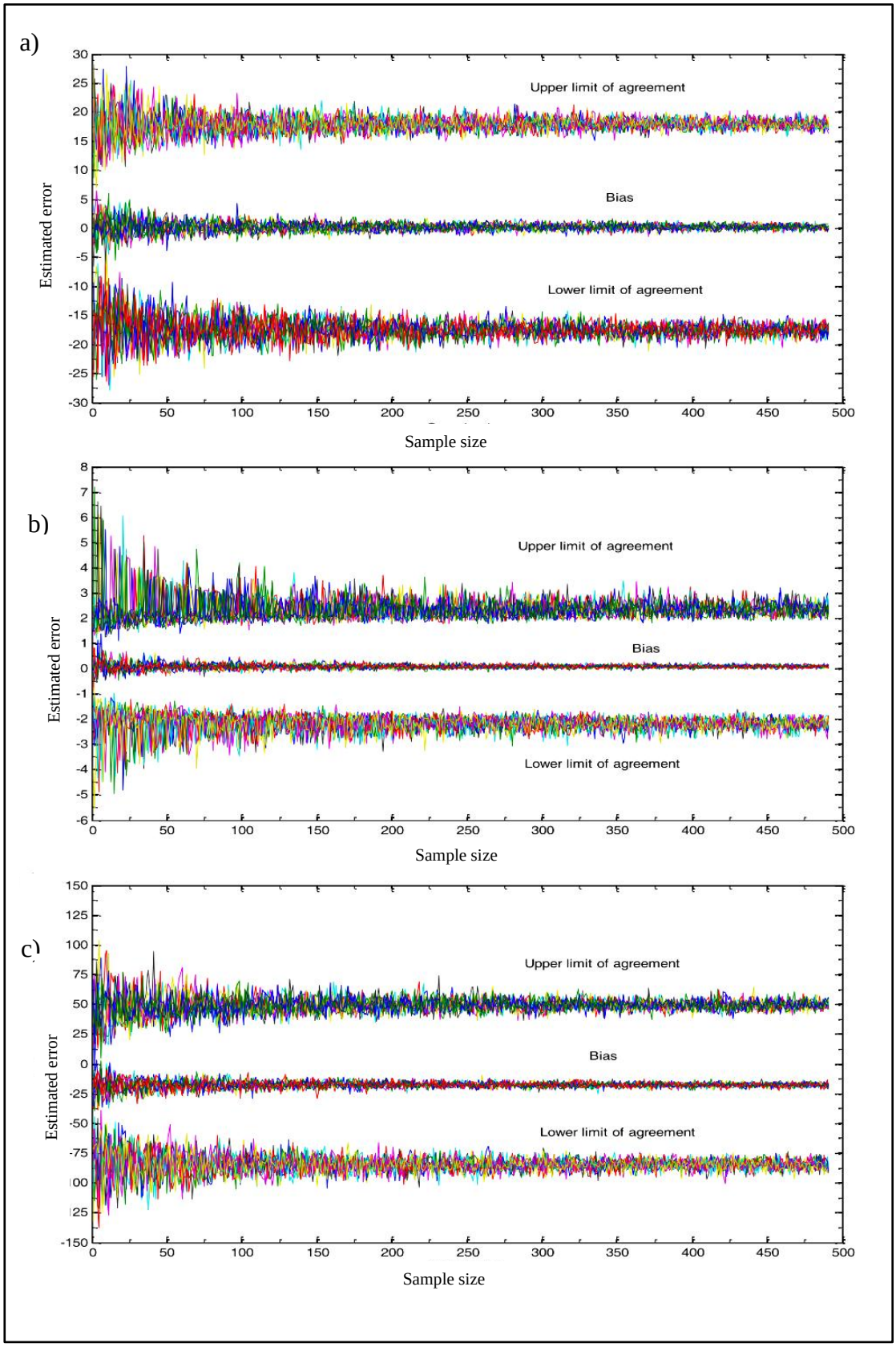


Figure 4.12: The plot of Bland-Altman limits of agreement versus sample size for all 10 sets of analysis for DBP (a), weight (b) and PEFR (c).

Figure 4.13 shows the result for the estimation of bias for glucose variable using ICC_A. The predictions become more stable when the sample size is more than 100. For instance, when the sample size is 25, the analysis of simulated data for blood glucose level shows that the prediction of ICC_A is between 0.81 and as high as 0.98. When the sample size is 50, the prediction of ICC_A is between 0.87 and 0.98. Similar trend were seen for systolic BP, diastolic BP, weight and PEFR. Figure 4.14 shows the result for systolic BP. Results for the other variables are shown in Figure 4.15.

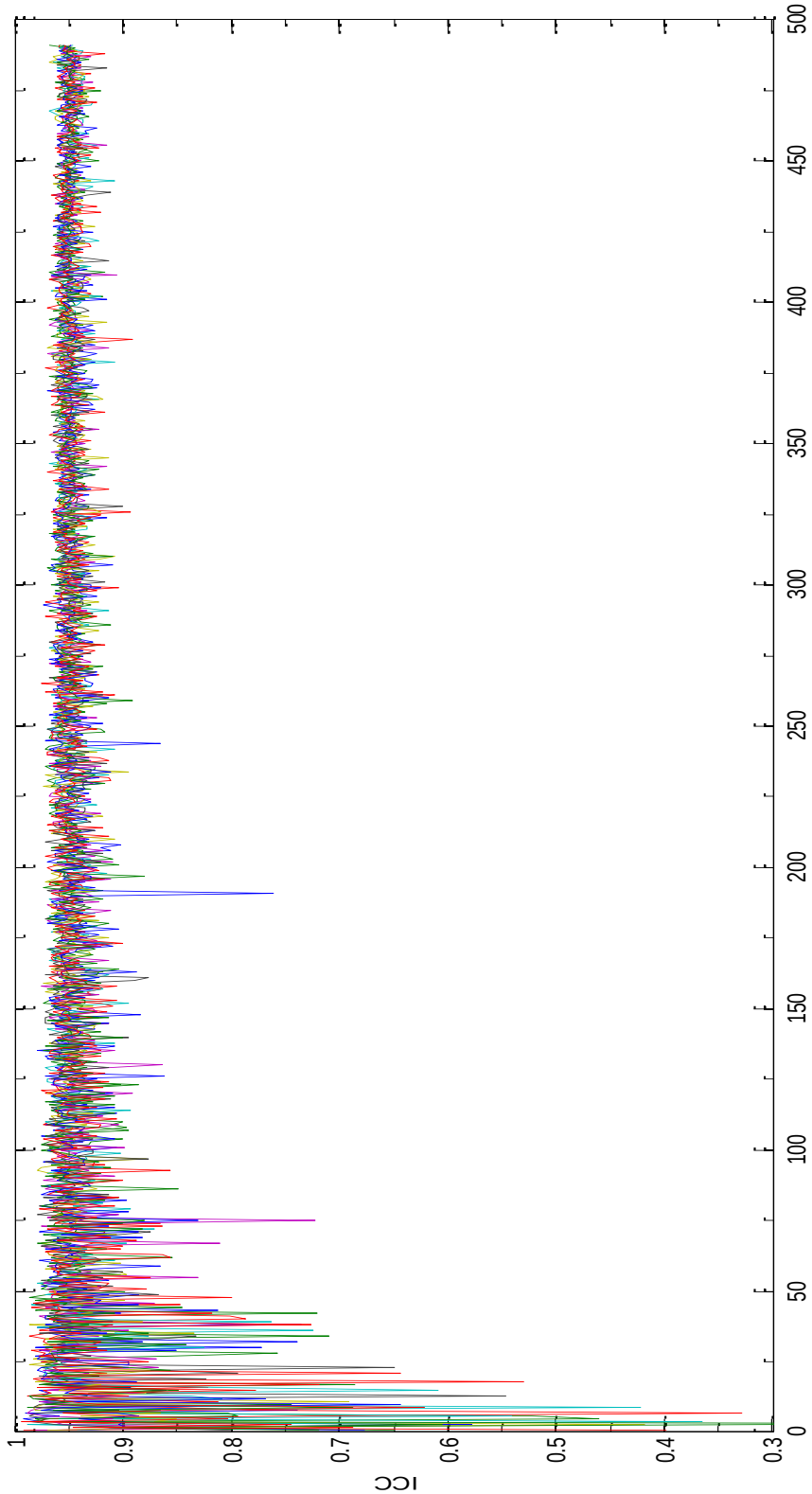


Figure 4.13: The plot of ICC_A versus sample size for all 10 sets of analysis (Glucose)

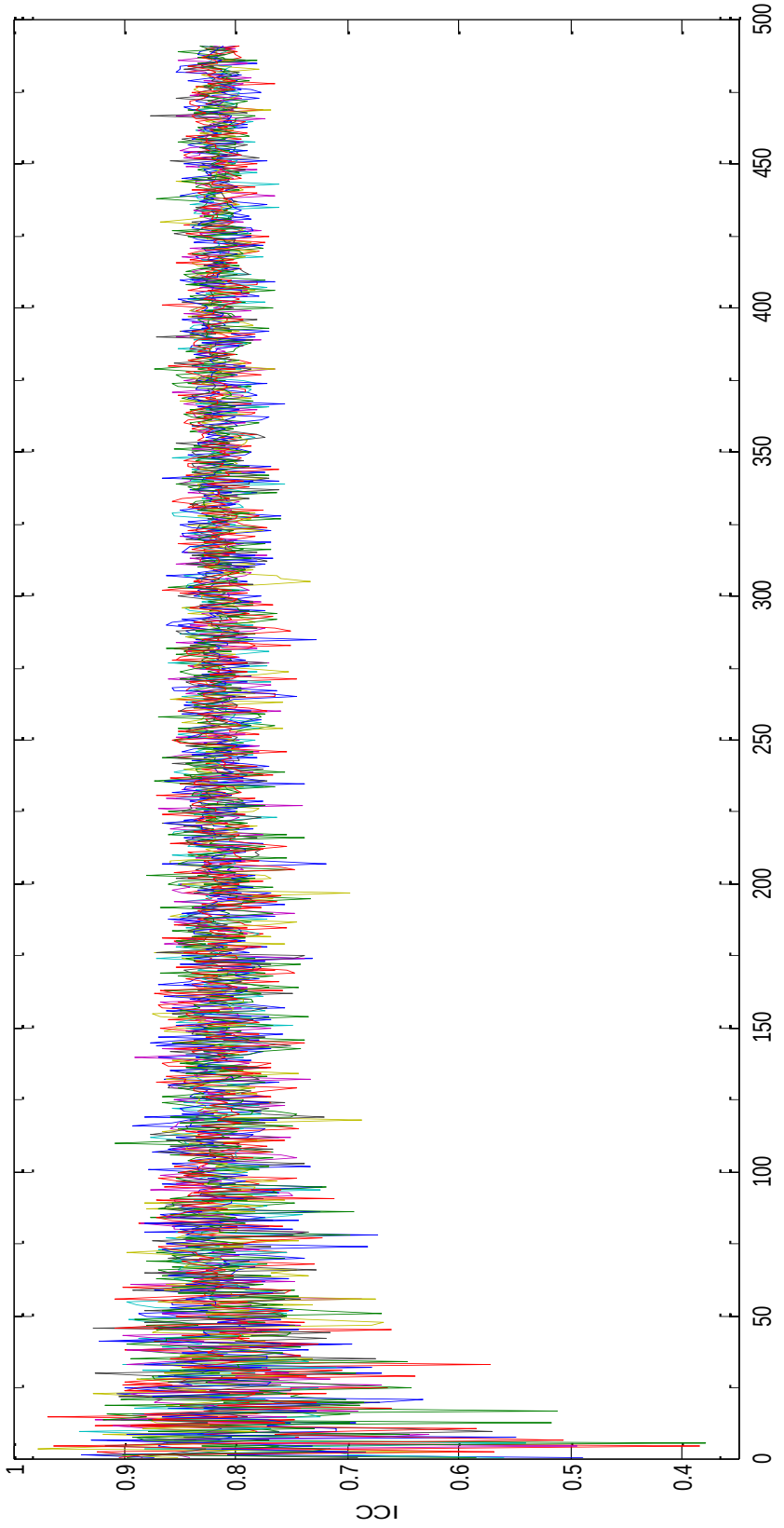


Figure 4.14: The plot of ICC_A versus sample size for all 10 sets of analysis (SBP)

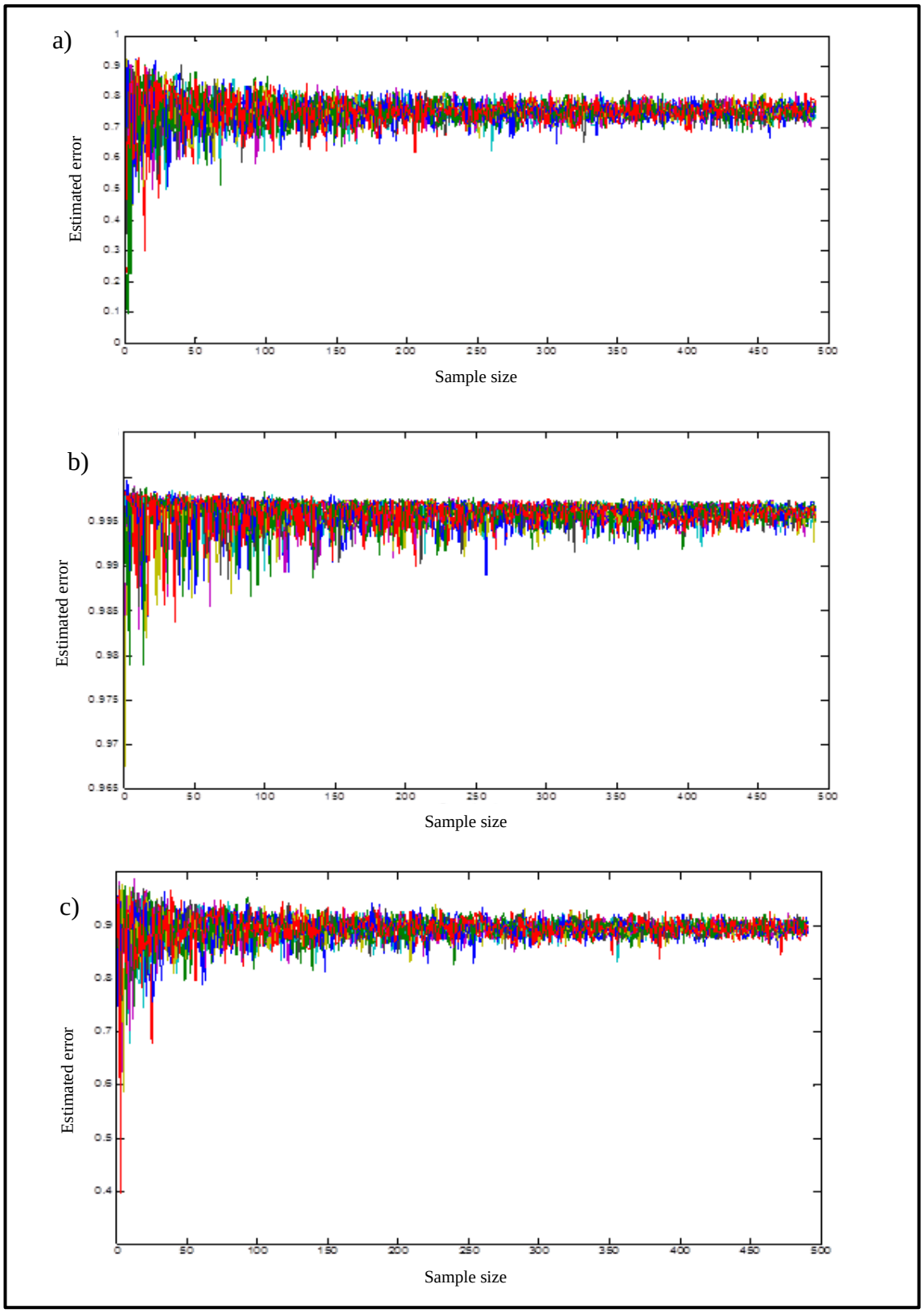


Figure 4.15: The plot of ICC_A versus sample size for all 10 sets of analysis for DBP (a), weight (b) and PEFR (c).

4.3.3 Extended analysis of the Bland-Altman method

Three variables (blood glucose level, body weight, and systolic BP) with sample sizes of 300 for each variable were analysed for this section. Twenty data sets with simulated bias were generated for each variable. The scatters of differences were uniform (homoscedasticity) for all data sets. The slopes of the regression lines of the Bland-Altman plot did not significantly differ from zero for all data sets. Therefore, proportional bias was assumed not to exist in these datasets.

Results for the analysis are summarised in Table 4.32 to Table 4.34. Figure 4.16 to Figure 4.18 shows the relationship of the range of predicted bias and the range of actual (simulated) bias. The pattern of relationship is similar for all three variables. All three graphs suggest that the range of simulated bias is higher than the actual bias. The overestimation of bias is increases as the range of bias increases.

Table 4.32: The prediction of generated bias and limits of agreement for the blood glucose level (mmol/l)

Simulated random error	Lower limit of agreement (CI)	Upper limit of agreement (CI)	Range of simulated error	Range of predicted error (LoA only)	Range of predicted error (LoA with CIs)
1. Error 0 to 0.1	-0.01 (-0.01 to 0)	0.11 (0.10 to 0.11)	0.1	0.12	0.12
2. Error 0 to 0.2	-0.01 (-0.02 to 0)	0.22 (0.20 to 0.23)	0.2	0.23	0.25
3. Error 0 to 0.3	-0.02 (-0.04 to -0.01)	0.32 (0.30 to 0.33)	0.3	0.34	0.37
4. Error 0 to 0.4	-0.04 (-0.06 to -0.01)	0.43 (0.41 to 0.46)	0.4	0.47	0.52
5. Error 0 to 0.5	-0.03 (-0.06 to -0.01)	0.52 (0.49 to 0.54)	0.5	0.55	0.6
6. Error 0 to 0.6	-0.02 (-0.05 to 0.01)	0.61 (0.58 to 0.64)	0.6	0.63	0.69
7. Error 0 to 0.7	-0.04 (-0.08 to 0)	0.75 (0.71 to 0.79)	0.7	0.79	0.87
8. Error 0 to 0.8	-0.05 (-0.09 to 0)	0.88 (0.84 to 0.93)	0.8	0.93	1.02
9. Error 0 to 0.9	-0.05 (-0.10 to 0)	0.96 (0.91 to 1.01)	0.9	1.01	1.11
10. Error 0 to 1.0	-0.06 (-0.12 to -0.01)	1.06 (1.01 to 1.12)	1.0	1.12	1.24
11. Error -0.1 to 0.1	-0.11 (-0.12 to -0.10)	0.12 (0.11 to 0.13)	0.2	0.23	0.25
12. Error -0.2 to 0.2	-0.22 (-0.24 to 0.20)	0.22 (0.20 to 0.24)	0.4	0.44	0.48
13. Error -0.3 to 0.3	-0.34 (-0.37 to -0.30)	0.33 (0.30 to 0.37)	0.6	0.67	0.74
14. Error -0.4 to 0.4	-0.49 (-0.53 to -0.44)	0.43 (0.38 to 0.47)	0.8	0.92	1.00
15. Error -0.5 to 0.5	-0.58 (-0.64 to -0.52)	0.58 (0.53 to 0.64)	1.0	1.16	1.28
16. Error -0.6 to 0.6	-0.61 (-0.67 to -0.54)	0.69 (0.62 to 0.75)	1.2	1.30	1.42
17. Error -0.7 to 0.7	-0.80 (-0.88 to -0.73)	0.74 (0.66 to 0.81)	1.4	1.54	1.69
18. Error -0.8 to 0.8	-0.92 (-1.01 to -0.83)	0.90 (0.81 to 0.99)	1.6	1.82	2.00
19. Error -0.9 to 0.9	-1.06 (-1.16 to -0.95)	1.05 (0.94 to 1.15)	1.8	2.11	2.31
20. Error -1.0 to 1.0	-1.06 (-1.16 to -0.94)	1.14 (1.03 to 1.24)	2.0	2.20	2.40

Table 4.33: The prediction of generated bias and limits of agreement for the body weight (kg)

Simulated random error	Lower limit of agreement (CI)	Upper limit of agreement (CI)	Range of simulated error	Range of predicted error (LoA only)	Range of predicted error (LoA with CIs)
1. Error 0 to 0.1	-0.01 (-0.01 to 0)	0.11 (0.10 to 0.11)	0.1	0.11	0.12
2. Error 0 to 0.2	-0.01 (-0.02 to 0)	0.21 (0.19 to 0.22)	0.2	0.22	0.24
3. Error 0 to 0.3	-0.03 (-0.05 to -0.01)	0.31 (0.30 to 0.31)	0.3	0.34	0.36
4. Error 0 to 0.4	-0.02 (-0.04 to 0.01)	0.43 (0.41 to 0.45)	0.4	0.45	0.49
5. Error 0 to 0.5	-0.01 (-0.04 to 0.02)	0.54 (0.51 to 0.56)	0.5	0.55	0.60
6. Error 0 to 0.6	-0.05 (-0.08 to -0.01)	0.63 (0.60 to 0.66)	0.6	0.68	0.74
7. Error 0 to 0.7	-0.03 (-0.07 to 0.01)	0.76 (0.72 to 0.80)	0.7	0.79	0.87
8. Error 0 to 0.8	-0.06 (-0.11 to -0.01)	0.86 (0.81 to 0.90)	0.8	0.92	1.01
9. Error 0 to 0.9	-0.05 (-0.10 to 0)	0.97 (0.92 to 1.02)	0.9	1.02	1.12
10. Error 0 to 1.0	-0.05 (-0.11 to 0)	1.08 (1.03 to 1.14)	1.0	1.13	1.25
11. Error -0.1 to 0.1	-0.11 (-0.12 to -0.10)	0.11 (0.10 to 0.12)	0.2	0.22	0.24
12. Error -0.2 to 0.2	-0.21 (-0.23 to 0.19)	0.23 (0.21 to 0.25)	0.4	0.44	0.48
13. Error -0.3 to 0.3	-0.32 (-0.36 to -0.29)	0.33 (0.30 to 0.37)	0.6	0.65	0.73
14. Error -0.4 to 0.4	-0.47 (-0.51 to -0.42)	0.48 (0.44 to 0.53)	0.8	0.95	1.04
15. Error -0.5 to 0.5	-0.59 (-0.65 to -0.53)	0.57 (0.51 to 0.63)	1.0	1.16	1.28
16. Error -0.6 to 0.6	-0.66 (-0.73 to -0.59)	0.70 (0.64 to 0.77)	1.2	1.36	1.50
17. Error -0.7 to 0.7	-0.80 (-0.87 to -0.72)	0.77 (0.70 to 0.85)	1.4	1.57	1.72
18. Error -0.8 to 0.8	-0.90 (-0.99 to -0.81)	0.88 (0.79 to 0.97)	1.6	1.78	1.96
19. Error -0.9 to 0.9	-1.06 (-1.16 to -0.96)	0.97 (0.87 to 1.07)	1.8	2.03	2.23
20. Error -1.0 to 1.0	-1.12 (-1.23 to -1.01)	1.07 (0.97 to 1.18)	2.0	2.19	2.41

Table 4.34: The prediction of generated bias and limits of agreement for the systolic BP (mmHg)

Simulated error	Lower limit of agreement (CI)	Upper limit of agreement (CI)	Range of simulated error	Range of predicted error (LoA only)	Range of predicted error (LoA with CIs)
1. Error 0 to 2	0 (-1 to 0)	2 (2 to 3)	2	2	4
2. Error 0 to 4	0 (-1 to 0)	4 (4 to 5)	4	4	6
3. Error 0 to 6	0 (-1 to 0)	7 (6 to 7)	6	7	8
4. Error 0 to 8	0 (-1 to 0)	9 (8 to 9)	8	9	10
5. Error 0 to 10	0 (-1 to 0)	11 (10 to 11)	10	11	12
6. Error 0 to 12	-1 (-2 to 0)	13 (12 to 14)	12	14	16
7. Error 0 to 14	-1 (-2 to 0)	14 (13 to 15)	14	15	17
8. Error 0 to 16	-1 (-2 to 0)	17 (16 to 18)	16	28	20
9. Error 0 to 18	-2 (-3 to -1)	19 (18 to 20)	18	21	23
10. Error 0 to 20	-1 (-2 to 0)	22 (20 to 23)	20	23	25
11. Error -2 to 2	-2 (-2 to 2)	2 (2 to 3)	4	4	5
12. Error -4 to 4	-4 (-5 to -4)	5 (4 to 5)	8	9	10
13. Error -6 to 6	-7 (-8 to -6)	7 (7 to 8)	12	14	16
14. Error -8 to 8	-10 (-10 to -9)	9 (8 to 10)	16	19	20
15. Error -10 to 10	-11 (-12 to -10)	11 (10 to 12)	20	22	24
16. Error -12 to 12	-13 (-15 to -12)	14 (12 to 15)	24	27	30
17. Error -14 to 14	-16 (-17 to -14)	17 (15 to 18)	28	33	35
18. Error -16 to 16	-19 (-21 to -17)	18 (16 to 20)	32	37	41
19. Error -18 to 18	-19 (-21 to -17)	21 (19 to 23)	36	40	44
20. Error -20 to 20	-28 (-31 to -25)	32 (29 to 35)	40	60	66

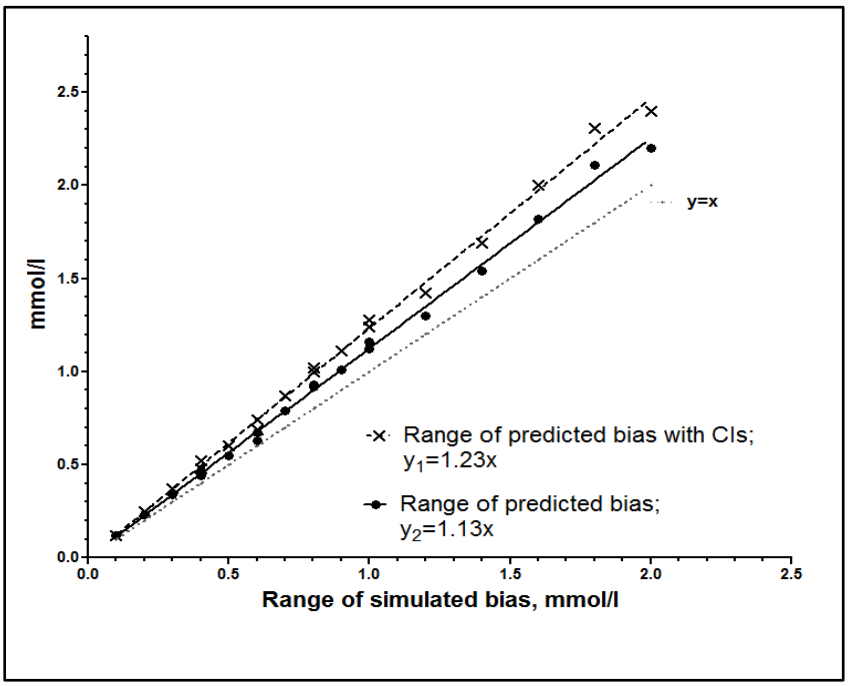


Figure 4.16: Relationship between the simulated and predicted error in the Bland-Altman analysis for blood glucose level

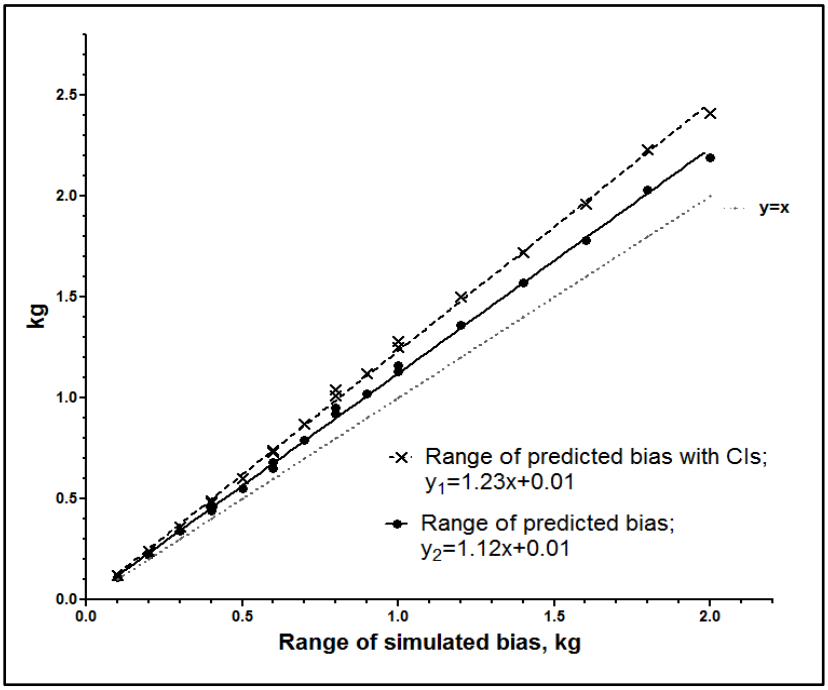


Figure 4.17: Relationship between the simulated and predicted error in the Bland-Altman analysis for body weight.

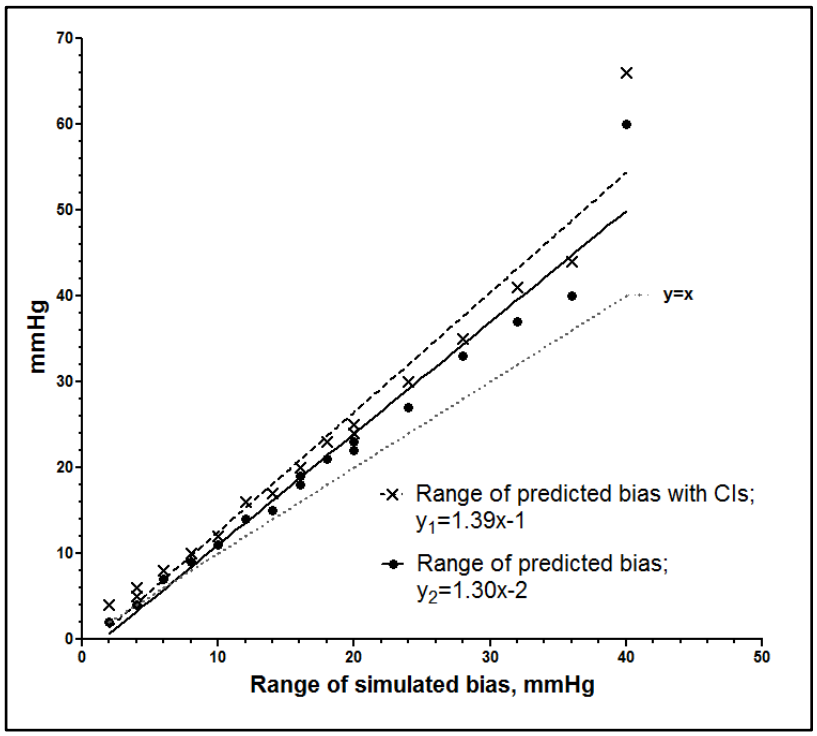


Figure 4.18: Relationship between the simulated and predicted error in the Bland-Altman analysis for systolic blood pressure.

4.4 Reliability Analysis

4.4.1 Prediction of Reliability: Clinical data

4.4.1.1 Comparison of Statistical Method in Analysis of Reliability

The Bland-Altman LoA, ICC_C, and ICC_A were conducted with samples of 300 for each variable (blood glucose level, systolic BP, diastolic BP, body temperature, carbon monoxide level and heart rate). The purpose of this analysis is to compare the ability of each method in predicting reliability of each instrument that measures all the variables. The interpretation of Bland-Altman method in the analysis of reliability was similar to the agreement analysis (as explained in Section 3.9.3.3). The value of ICC $\geq$ 0.75 was considered to be good reliability (Rosner, 2006).

Data used for the analysis were collected in clinical settings, and all instruments were validated by their manufacturer. Therefore, the instruments should be reliable. The ICC_C and ICC_A predict that all instruments were reliable. However, Bland-Altman LoA only shows good reliability for infrared thermometer (temperature). All other instruments were found to have poor reliability by the Bland-Altman method. Results for this analysis are summarised in Table 4.35.

Table 4.35: Prediction of reliability for all variables using clinical data

Variables (n=300)	Bland-Altman LoA	ICC_C	ICC_A
SBP	-3mmHg (-20 to -15) Poor Reliability	0.902 (0.883 to 0.919) Good Reliability	0.897 (0.873 to 0.917) Good Reliability
DBP	-1mmHg (-14 to 11) Poor Reliability	0.882 (0.859 to 0.902) Good Reliability	0.875 (0.846 to 0.899) Good Reliability
Temperature	-0.01^0C (-0.17 to 0.16) Good Reliability	0.966 (0.959 to 0.972) Good Reliability	0.966 (0.959 to 0.972) Good Reliability
PEFR	13l/min (-70 to 96) Poor Reliability	0.879 (0.856 to 0.899) Good Reliability	0.864 (0.820 to 0.896) Good Reliability
CO level	0ppm (-5 to 5) Poor Reliability	0.979 (0.975 to 0.983) Good Reliability	0.940 (0.928 to 0.951) Good Reliability
Heart rate	0bpm (-3 to 3) Poor Reliability	0.987 (0.985 to 0.990) Good Reliability	0.987 (0.985 to 0.990) Good Reliability

4.4.1.2 Consistency of Prediction

For this analysis, 10 sets of data with sample of 200 were selected randomly from the total of 300 samples for each variable. The purpose of this analysis is to compare the consistency of each statistical method in predicting reliability of instruments measuring each variable. All instruments were validated by their manufacturer at the beginning of this study, therefore all instrument assumed to be reliable and should produce consistent result. Results for this analysis are summarised in Table 4.36 to Table 4.41.

All three statistical methods provide a consistent prediction of reliability for all ten sets of data for all variables. The summaries of result for the consistency of prediction of reliability for all methods are displayed in Table 4.42.

Table 4.36: Prediction of reliability of instrument measuring SBP

Set	Bland-Altman LoA	ICC_C	ICC_A
1	-2mmHg (-18 to 14) Poor Reliability	0.915 (0.894 to 0.973) Good Reliability	0.911 (0.887 to 0.931) Good Reliability
2	-3mmHg (-20 to 15) Poor Reliability	0.902 (0.878 to 0.923) Good Reliability	0.896 (0.866 to 0.920) Good Reliability
3	-2mmHg (-19 to 15 Poor Reliability	0.892 (0.865 to 0.914) Good Reliability	0.888 (0.860 to0.912) Good Reliability
4	-2mmHg (-18 to 14) Poor Reliability	0.909 (0.886 to 0.928) Good Reliability	0.902 (0.872 to 0.925) Good Reliability
5	-2mmHg (-20 to 16) Poor Reliability	0.902 (0.878 to 0.922) Good Reliability	0.898 (0.871 to 0.920) Good Reliability
6	-3mmHg (-19 to 14 Poor Reliability	0.906 (0.883 to 0.925) Good Reliability	0.899 (0.870 to 0.922) Good Reliability
7	-3mmHg (-19 to 14) Poor Reliability	0.917 (0.897 to 0.934) Good Reliability	0.912 (0.887 to 0.932) Good Reliability
8	-2mmHg (-19 to 15) Poor Reliability	0.894 (0.868 to 0.916) Good Reliability	0.890 (0.862 to 0.914) Good Reliability
9	-3mmHg (-20 to 14) Poor Reliability	0.915 (0.894 to 0.932) Good Reliability	0.908 (0.881 to 0.930) Good Reliability
10	-3mmHg (-1 to 10) Poor Reliability	0.906 (0.883 to 0.926) Good Reliability	0.899 (0.870 to 0.923) Good Reliability

Table 4.37: Prediction of reliability of instrument measuring DBP

Set	Bland-Altman LoA	ICC_C	ICC_A
1	-1mmHg (-15 to 12) Poor Reliability	0.892 (0.866 to 0.914) Good Reliability	0.886 (0.856 to 0.911) Good Reliability
2	-2mmHg (-13 to 10) Poor Reliability	0.879 (0.850 to 0.903) Good Reliability	0.869 (0.830 to 0.899) Good Reliability
3	-1mmHg (-14 to 12) Poor Reliability	0.868 (0.837 to 0.895) Good Reliability	0.861 (0.824 to 0.891) Good Reliability
4	-2mmHg (-12 to 9) Poor Reliability	0.913 (0.892 to 0.931) Good Reliability	0.905 (0.875 to 0.928) Good Reliability
5	-2mmHg (-15 to 11) Poor Reliability	0.885 (0.858 to 0.909) Good Reliability	0.874 (0.833 to 0.905) Good Reliability
6	-1mmHg (-15 to 12) Poor Reliability	0.863 (0.831 to 0.891) Good Reliability	0.857 (0.821 to 0.888) Good Reliability
7	-1mmHg (-13 to 12 Poor Reliability	0.881 (0.853 to 0.906) Good Reliability	0.874 (0.840 to 0.902) Good Reliability
8	-1mmHg (-12 to 10) Poor Reliability	0.913 (0.892 to 0.931) Good Reliability	0.908 (0.882 to 0.929) Good Reliability
9	-2mmHg (-15 to 11) Poor Reliability	0.885 (0.857 to 0.908) Good Reliability	0.873 (0.833 to 0.904) Good Reliability
10	-1mmHg (-14 to 11) Poor Reliability	0.892 (0.866 to 0.915) Good Reliability	0.887 (0.857 to 0.912) Good Reliability

Table 4.38**:** Prediction of reliability of instrument measuring temperature

Set	Bland-Altman LoA	ICC_C	ICC_A
1	-0.01^{0}C (-0.18 to 0.16) Good Reliability	0.970 (0.963 to 0.977) Good Reliability	0.970 (0.962 to 0.977) Good Reliability
2	-0.01^{0}C (-0.16 to 0.14) Good Reliability	0.976 (0.973 to 0.983) Good Reliability	0.979 (0.973 to 0.983) Good Reliability
3	-0.00^{0}C (-0.15 to 0.15) Good Reliability	0.962 (0.952 to 0.970) Good Reliability	0.962 (0.952 to 0.970) Good Reliability
4	-0.01^{0}C (-0.14 to 0.13) Good Reliability	0.966 (0.957 to 0.973) Good Reliability	0.966 (0.957 to 0.973) Good Reliability
5	-0.01^{0}C (-0.15 to 0.15) Good Reliability	0.974 (0.967 to 0.979) Good Reliability	0.974 (0.967 to 0.979) Good Reliability
6	-0.00^{0}C (-0.16 to 0.15) Good Reliability	0.975 (0.969 to 0.981) Good Reliability	0.975 (0.969 to 0.981) Good Reliability
7	-0.01^{0}C (-0.16 to 0.15) Good Reliability	0.976 (0.970 to 0.981) Good Reliability	0.976 (0.970 to 0.981) Good Reliability
8	-0.01^{0}C (-0.17 to 0.16) Good Reliability	0.962 (0.952 to 0.970) Good Reliability	0.962 (0.952 to 0.970) Good Reliability
9	-0.01^{0}C (-0.18 to 0.17) Good Reliability	0.957 (0.946 to 0.966) Good Reliability	0.957 (0.946 to 0.966) Good Reliability
10	-0.00^{0}C (-0.16 to 0.15) Good Reliability	0.973 (0.966 to 0.979) Good Reliability	0.973 (0.966 to 0.979) Good Reliability

Table 4.39: Prediction of reliability of instrument measuring PEFR

Set	Bland-Altman LoA	ICC_C	ICC_A
1	14l/min (-78 to 105) Poor Reliability	0.871 (0.840 to 0.897) Good Reliability	0.857 (0.809 to 0.892) Good Reliability
2	14 l/min (-72 to 101) Poor Reliability	0.870 (0.839 to 0.896) Good Reliability	0.852 (0.798 to 0.891) Good Reliability
3	12 l/min (-75 to 99) Poor Reliability	0.874 (0.844 to 0.900) Good Reliability	0.860 (0.812 to 0.895) Good Reliability
4	13 l/min (-75 to 100) Poor Reliability	0.878 (0.848 to 0.903) Good Reliability	0.862 (0.812 to 0.898) Good Reliability
5	14 l/min (-62 to 89) Poor Reliability	0.889 (0.862 to 0.912) Good Reliability	0.875 (0.830 to 0.908) Good Reliability
6	15 l/min (-67 to 96) Poor Reliability	0.871 (0.840 to 0.897) Good Reliability	0.856 (0.807 to 0.893) Good Reliability
7	13 l/min (-71 to 98) Poor Reliability	0.875 (0.845 to 0.901) Good Reliability	0.861 (0.813 to 0.896) Good Reliability
8	13 l/min (-68 to 94) Poor Reliability	0.877 (0.847 to 0.902) Good Reliability	0.859 (0.804 to 0.897) Good Reliability
9	14 l/min (-72 to 101) Poor Reliability	0.881 (0.852 to 0.905) Good Reliability	0.864 (0.812 to 0.901) Good Reliability
10	15 l/min (-65 to 95) Poor Reliability	0.875 (0.845 to 0.901) Good Reliability	0.861 (0.814 to 0.896) Good Reliability

Table 4.40: Prediction of reliability of instrument measuring CO level

Set	Bland-Altman LoA	ICC_C	ICC_A
1	0ppm (-5 to 5) Poor Reliability	0.935 (0.919 to 0.949) Good Reliability	0.935 (0.919 to 0.949) Good Reliability
2	0ppm (-5 to 5) Poor Reliability	0.948 (0.935 to 0.959) Good Reliability	0.948 (0.935 to 0.959) Good Reliability
3	0ppm (-4 to 4) Poor Reliability	0.940 (0.925 to 0.953) Good Reliability	0.940 (0.925 to 0.953) Good Reliability
4	0ppm (-5 to 5) Poor Reliability	0.934 (0.918 to 0.948) Good Reliability	0.935 (0.918 to 0.949) Good Reliability
5	0ppm (-5 to 5) Poor Reliability	0.931 (0.913 to 0.945) Good Reliability	0.931 (0.914 to 0.945) Good Reliability
6	0ppm (-5 to 5) Poor Reliability	0.937 (0.922 to 0.951) Good Reliability	0.938 (0.922 to 0.951) Good Reliability
7	0ppm (-4 to 4) Poor Reliability	0.939 (0.923 to 0.952) Good Reliability	0.939 (0.924 to 0.952) Good Reliability
8	0ppm (-5 to 5) Poor Reliability	0.942 (0.928 to 0.954) Good Reliability	0.942 (0.928 to 0.954) Good Reliability
9	0ppm (-5 to 5) Poor Reliability	0.936 (0.919 to 0.949) Good Reliability	0.936 (0.919 to 0.949) Good Reliability
10	0ppm (-4 to 4) Poor Reliability	0.942 (0.928 to 0.985) Good Reliability	0.943 (0.928 to 0.955) Good Reliability

Table 4.41: Prediction of reliability of instrument measuring Heart rate

Set	Bland-Altman LoA	ICC_C	ICC_A
1	0bpm (-3 to 3) Poor Reliability	0.987 (0.984 to 0.990) Good Reliability	0.987 (0.983 to 0.990) Good Reliability
2	0bpm (-3 to 3) Poor Reliability	0.989 (0.986 to 0.991) Good Reliability	0.989 (0.986 to 0.991) Good Reliability
3	0bpm (-3 to 3) Poor Reliability	0.987 (0.984 to 0.990) Good Reliability	0.987 (0.983 to 0.990) Good Reliability
4	0bpm (-3 to 3) Poor Reliability	0.988 (0.985 to 0.990) Good Reliability	0.988 (0.985 to 0.990) Good Reliability
5	0bpm (-3 to 3) Poor Reliability	0.988 (0.985 to 0.991) Good Reliability	0.988 (0.985 to 0.991) Good Reliability
6	0bpm (-3 to 3) Poor Reliability	0.988 (0.985 to 0.991) Good Reliability	0.988 (0.985 to 0.991) Good Reliability
7	0bpm (-3 to 3) Poor Reliability	0.986 (0.982 to 0.989) Good Reliability	0.986 (0.982 to 0.989) Good Reliability
8	0bpm (-3 to 3) Poor Reliability	0.988 (0.985 to 0.991) Good Reliability	0.988 (0.985 to 0.991) Good Reliability
9	0bpm (-3 to 3) Poor Reliability	0.986 (0.983 to 0.989) Good Reliability	0.986 (0.982 to 0.989) Good Reliability
10	0bpm (-3 to 3) Poor Reliability	0.986 (0.982 to 0.989) Good Reliability	0.986 (0.982 to 0.989) Good Reliability

Table 4.42: Summary of prediction of reliability for all 10 clinical data set.

Statistic method	SBP	DBP	Temperature	CO	PEFR	HR
1. Bland-Altman LoA	NO -10	NO -10	YES -10	NO -10	NO -10	NO -10
2. ICC_C	YES - 10	YES - 10	YES - 10	YES - 10	YES - 10	YES - 10
3. ICC_A	YES - 10	YES - 10	YES – 10	YES - 10	YES - 10	YES - 10

NO – Poor/moderate reliability of instrument measuring the variable.
YES - Good reliability of instrument measuring the variable.

4.4.1.3 Number of Measurement

The ICC_C, and ICC_A analysis were conducted with two and three sets of measurements. The purpose of this analysis is to compare the prediction of reliability of each instrument with different number of measurement. Both ICC_C and ICC_A provide similar prediction of reliability with two and three measurements. Results for this analysis are summarised in Table 4.43.

Table 4.43: Prediction of reliability with different number of measurement

Variables	ICC_C with 3 sets of reading (95% CI)	ICC_C with 2 sets of reading (95% CI)	ICC_A with 2 sets of reading (95% CI)	ICC_A with 3 sets of reading (95% CI)
SBP	0.957(0.948 to 0.965) Good Reliability	0.937(0.921 to 0.950) Good Reliability	0.897(0.862 to 0.921) Good Reliability	0.897 (0.873 to 0.917) Good Reliability
DBP	0.965(0.958 to 0.972) Good Reliability	0.949(0.936 to 0.960) Good Reliability	0.877(0.845 to 0.902) Good Reliability	0.875 (0.864 to 0.899) Good Reliability
Temperature	0.988(0.986 to 0.990) Good Reliability	0.987(0.984 to 0.990) Good Reliability	0.975 (0.969 to 0.980) Good Reliability	0.966 (0.959 to 0.972) Good Reliability
PEFR	0.956(0.947 to 0.964) Good Reliability	0.934(0.917 to 0.947) Good Reliability	0.866(0.818 to 0.899) Good Reliability	0.864(0.820 to 0.896) Good Reliability
CO level	0.979(0.975 to 0.983) Good Reliability	0.964(0.955 to 0.971) Good Reliability	0.930 (0.913 to 0.944) Good Reliability	0.940 (0.928 to 0.951) Good Reliability
Heart rate	0.966(0.995 to 0.997) Good Reliability	0.995(0.994 to 0.996) Good Reliability	0.990 (0.987 to 0.992) Good Reliability	0.987(0.985 to 0.990) Good Reliability

4.4.2 Prediction of Reliability: Simulated data

In this analysis, five sets of simulated data were produced for each variable. The characteristic of the data set were as follow:

Set 1 (first measurement): original clinical data

Set 2 (second measurement): 1/3 constant positive error

Set 3 (third measurement): constant negative error

Set 4 (fourth measurement): 1/2 constant positive

Set 5 (fifth measurement): constant positive error

All simulated data set were purposely generated to represent unreliable measurement of instrument. It is clear that the five measurements (from set 1 to set 5) are not repeatable (i.e. the instrument producing this measurements is not reliable). For the systolic blood pressure variable, the ICC_C did not predict the poor reliability with all analysis of different number of measurements. The ICC_A predicts the poor reliability in all the analysis except the analysis with two sets of measurement. Results are shown in Table 4.44.

Table 4.44: Prediction of reliability with different number of measurements for SBP

	ICC_C	**ICC_A**
Five measurements (Set 1,2,3,4&5)	0.979 (0.975 to 0.983) Good Reliability	0.854 (0.501 to 0.938) Poor Reliability
Four measurements (Set 1,2,3&4)	0.977 (0.972 to 0.981) Good Reliability	0.876 (0.484 to 0.951) Poor Reliability
Three measurements (Set 1,2&3)	0.980 (0.975 to 0.983) Good Reliability	0.866 (0.246 to 0.955) Poor Reliability
Two measurements (Set 1&2)	0.971 (0.964 to 0.977) Good Reliability	0.958 (0.869 to 0.980) Good Reliability

For the diastolic blood pressure variable, the ICC_C did not predict the poor reliability with all analysis of different number of measurements. The ICC_A predicts the poor reliability in all the analysis except the analysis with two sets of measurement. Results are shown in the Table 4.45.

Table 4.45: Prediction of reliability with different number of measurements for DBP

	ICC_C	**ICC_A**
Five measurements (Set 1,2,3,4&5)	0.941 (0.931 to 0.951) Good Reliability	0.664 (0.250 to 0.836) Poor Reliability
Four measurements (Set 1,2,3&4)	0.930 (0.917 to 0.941) Good Reliability	0.689 (0.219 to 0.859) Poor Reliability
Three measurements (Set 1,2&3)	0.946 (0.935 to 0.955) Good Reliability	0.700 (0.090 to 0.885) Poor Reliability
Two measurements (Set 1&2)	0.915 (0.894 to 0.931) Good Reliability	0.878 (0.670 to 0.939) Moderate Reliability

For the body temperature variable, the ICC_C did not predict the poor reliability with all analysis of different number of measurements. The ICC_A predicts the poor or moderate reliability (i.e. not reliable) in all the analysis. Results are shown in Table 4.46.

Table 4.46: Prediction of reliability with different number of measurements for Temperature

	ICC_C	**ICC_A**
Five measurements (Set 1,2,3,4&5)	0.896 (0.878 to 0.912) Good Reliability	0.514 (0.148 to 0.731) Poor Reliability
Four measurements (Set 1,2,3&4)	0.880 (0.859 to 0.899) Good Reliability	0.551 (0.128 to 0.771) Poor Reliability
Three measurements (Set 1,2&3)	0.899 (0.879 to 0.916) Good Reliability	0.543 (0.035 to 0.797) Poor Reliability
Two measurements (Set 1&2)	0.850 (0.815 to 0.878) Good Reliability	0.791 (0.509 to 0.891) Moderate Reliability

For the peak expiratory flow rate variable, the ICC_C did not predict poor reliability with all analysis of different number of measurements. The ICC_A predicts moderate or poor reliability (i.e. not reliable) in all the analysis except the analysis with two sets of measurement. Results are shown in the Table 4.47.

Table 4.47: Prediction of reliability with different number of measurements for PEFR

	ICC_C	**ICC_A**
Five measurements (Set 1,2,3,4&5)	0.982 (0.979 to 0.985) Good Reliability	0.870 (0.536 to 0.945) Moderate Reliability
Four measurements (Set 1,2,3&4)	0.979 (0.975 to 0.983) Good Reliability	0.888 (0.513 to 0.956) Moderate Reliability
Three measurements (Set 1,2&3)	0.983 (0.979 to 0.986) Good Reliability	0.884 (0.280 to 0.962) Poor Reliability
Two measurements (Set 1&2)	0.975 (0.968 to 0.980) Good Reliability	0.962 (0.882 to 0.982) Good Reliability

For the carbon monoxide level variable, the ICC_C did not predict the poor reliability with all analysis of different number of measurements. The ICC_A predicts the poor reliability in all the analysis except the analysis with two sets of measurement. Results are shown in the Table 4.48.

Table 4.48: Prediction of reliability with different number of measurements for CO level

	ICC_C	**ICC_A**
Five measurements (Set 1,2,3,4&5)	0.964 (0.958 to 0.970) Good Reliability	0.769 (0.363 to 0.896) Poor Reliability
Four measurements (Set 1,2,3&4)	0.960 (0.952 to 0.967) Good Reliability	0.800 (0.343 to 0.917) Poor Reliability
Three measurements (Set 1,2&3)	0.965 (0.958 to 0.971) Good Reliability	0.786 (0.147 to 0.924) Poor Reliability
Two measurements (Set 1&2)	0.950 (0.937 to 0.960) Good Reliability	0.927 (0.784 to 0.965) Good Reliability

For the heart rate variable, the ICC_C did not predict the poor reliability with all numbers of measurements. The ICC_A predicts the moderate or poor reliability (i.e. not reliable) in all the analyses. Results are shown in the Table 4.49.

Table 4.49: Prediction of reliability with different number of measurements for Heart rate

	ICC_C	**ICC_A**
Five measurements (Set 1,2,3,4&5)	0.939 (0.929 to 0.949) Good Reliability	0.657 (0.244 to 0.831) Poor Reliability
Four measurements (Set 1,2,3&4)	0.931 (0.918 to 0.942) Good Reliability	0.692 (0.222 to 0.861) Poor Reliability
Three measurements (Set 1,2&3)	0.942 (0.930 to 0.952) Good Reliability	0.685 (0.082 to 0.878) Poor Reliability
Two measurements (Set 1&2)	0.915 (0.895 to 0.932) Good Reliability	0.878 (0.672 to 0.940) Moderate Reliability

4.4.3 Extended analysis of Intra-class Correlation Coefficient

4.4.3.1 Comparison of two repeated reading and three repeated readings

a. Clinical data

The differences in the prediction of ICCs for two and three repeated measurements were not significant for both ICC_A and ICC_C. Details of the results are shown in the Table 4.50. The power of the analyses were calculated using OpenEpi Version 2 online epidemiological calculator ("OpenEpi, Version 2, open source calculator--PowerMean,"). To detect the differences of 0.05 between the predicted ICCs, the analysis of ICC_A has a power of 99.99%, and the analysis of ICC_C has a power of 100%. However, both analyses were not powerful enough to detect differences of 0.01 (Power = 20.64% for ICC_A and 23.24% for ICC_C).

Table 4.50: Comparison of the prediction of ICC with two and three repeated measurements for clinical data

Two repeated readings compared with three repeated readings	Paired Differences					
	Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
		Lower	Upper			
ICC_A	-0.000467	-0.002459	0.001526	-0.469	59	0.641
ICC_C	-0.002267	-0.004904	0.000371	-1.720	59	0.091

b. Simulated data

The differences on the prediction of ICCs for two and three repeated measurements were significant for both ICC_A and ICC_C. Details of the results are shown in the Table 4.51.

Table 4.51: Comparison of the prediction of ICC with two and three repeated measurements for simulated data

Two repeated readings compared with three repeated readings	Paired Differences					
	Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
		Lower	Upper			
ICC_C	0.0659167	0.0003369	0.1314964	2.212	11	0.049
ICC_A	0.0659167	0.0003369	0.1314964	2.212	11	0.049

4.5 Summary of Chapter 4

This chapter deals with the results of this study. The variables were collected from both the UM wellness health screening population and in the community. A total of 300 samples were collected per variable. The data covers a good range of normal values of all variables. Some of the variables have a range of abnormal values in the data. This includes blood glucose level, systolic blood pressure, diastolic blood pressure, weight, peak expiratory flow rate, and carbon monoxide level. However, the data does not represent the extreme abnormal values (i.e. very high or very low values) for all the variables.

4.5.1 Agreement Analysis

The proposed method of measuring agreement has two main analysis, these are the comparison of slopes and y-intercepts analysis and the agreement model. The comparison of slopes and y-intercepts analysis involve the comparison of slopes and intercepts of the regression line with the line of agreement (y = x). The agreement model was based on the function of error.

Section 4.3.1 dealt with the analysis of agreement using clinical data. Total of 55 sets of clinical data were analysed using each method. Since all instruments were validated by their manufacturer, they should be in agreement with the standard. However, for the analysis of agreement using clinical data, the comparison of slopes and y-intercepts analysis only shows agreement for instruments measuring systolic blood pressure and weight, and not for other variables (glucose, diastolic blood pressure, and peak expiratory flow rate). The Bland-Altman LoA incorrectly shows disagreement for all instruments. The agreement model and ICC_A correctly predict the agreement of all instruments. The agreement model, Bland-Altman LoA and ICC_A

provide a very consistent prediction of agreement for all the variables. The comparison of slopes and y-intercepts analysis only provides consistent prediction of agreement for weight and PEFR.

Section 4.3.2 dealt with the analysis of agreement using simulated data. The simulated data sets were designed to represent disagreement of instruments for all variables. The analyses test the ability of each method in predicting constant error, inconsistent error, and the effect of proportion of error in data set. Total of 45 sets of simulated data were analysed for each method.

All four methods correctly predict the constant error in the data set of all variables. The agreement model and the Bland-Altman LoA predict the simulated error in all the data sets. The y-intercept of linear regression also seems to reflect the error in all the data set. The actual value of ICC_A did not provide good prediction of agreement, but the confidence interval (CI) reflects the disagreement in the data set. The information from ICC_A analysis also does not provide information on the direction of error (positive or negative error). The patterns of prediction for all four methods were similar for both positive and negative error.

All methods did not provide good prediction when the errors in the data sets were not consistent. The comparison of slopes and y-intercepts analysis, agreement model and ICC_A detect the disagreement for some of the variables only. The Bland-Altman LoA shows the disagreement of data set with inconsistent error for all the variables. Although Bland-Altman LoA successfully detects the disagreement in all data sets, the actual biases predicted were overestimated. The overestimations of bias were seen for all variables.

The comparison of slopes and y-intercepts analysis and the Bland-Altman LoA show the disagreement in different data sets of various proportion of error for all the variables. The proportion of error influences the prediction of quantification and

direction of error in the data set. The proportion of error in the data set also has an effect on prediction of agreement by the analysis of error and ICC_A.

Section 4.3.2.4 tests the effect of sample size. The prediction of outcome for all methods (slope, intercept, predicted error using the agreement model, estimated bias using the Bland-Altman method, ULA, LLA, and ICC_A) becomes stable as the sample size increases. The standard error of all outcomes for each method decreases as the sample size increases, and stabilise after a certain sample size. Most of the prediction of the outcomes for all methods stabilise after the sample size is more than 100. The pattern of prediction become more constant after the sample size is greater than 200. The pattern is similar for the analysis of various ranges of variables (small or wide range of variables).

The extended analysis of the Bland-Altman method (Section 4.3.3) shows that there is an overestimation in the prediction of bias in the Bland-Altman analysis even when proportional bias had been shown to be absent by testing the slope of the regression line in the Bland-Altman plot. The overestimation of bias increases when the range of actual bias increases. Similar pattern were seen for all three variables which means this is not an isolated issue.

4.5.5 Reliability Analysis

Section 4.4 deals with the analysis of reliability. The first part of analysis (section 4.4.1) was performed using clinical data. Total of 66 sets of clinical data were analysed using each method. All instruments were assumed to be reliable because all were validated by their manufacturer. The ICC_C and ICC_A predict that all instruments were reliable. However, the Bland-Altman LoA only shows good reliability for infrared thermometer (temperature). All other instruments were found to have poor reliability by the Bland-Altman method. All three methods provide a consistent prediction of reliability for all

ten sets of data for all variables. Both ICC_C and ICC_A provide similar prediction of reliability with two and three number of measurements.

The second part of reliability analysis (section 4.4.2) involves prediction of reliability of simulated data. Simulated data sets were generated to represent imprecision or unreliable instrument. The ICC_A provide better prediction of reliability compared to the ICC_C. The ICC_C did not predict the poor reliability with all analysis of different number of measurements. The ICC_A predicts the poor reliability in all the analysis except the analysis with two sets of measurement.

The next section (section 4.4.3) is the extended analysis of the Intra-class Correlation Coefficient. This section compares the prediction of ICCs (for both ICC_A and ICC_C) with two and three repeated measurements. For the clinical data, the paired t-test showed that there was no significant difference between the prediction of ICC_A with two and three repeated measurements. Similar result was seen for the prediction of ICC_C. However, the analysis of simulated data shows that there were significant differences between the prediction of ICCs (for both ICC_A and ICC_C) with two and three repeated measurements.

CHAPTER 5: DISCUSSION

5.1 Introduction

Agreement and reliability are both important parameters in determining the quality of an instrument. This study was designed to evaluate the statistical methods used to assess the agreement and reliability of medical instruments that measure continuous outcomes. This chapter will discuss the findings of this study.

This chapter will begin with a discussion of the results for the analysis of agreement in Section 5.2. The four methods tested for the agreement analysis were comparing slopes and y-intercepts analysis, agreement model, Bland-Altman Limits of Agreement, and Intra-class Correlation Coefficient for agreement (ICC_A).

Section 5.3 is a discussion of the results for the analysis of reliability. Statistical methods tested for the reliability analyses were the Bland-Altman Limits of Agreement (LoA), the Intra-class Correlation Coefficient for reliability (ICC_C), and the Intra-class Correlation Coefficient for agreement (ICC_A). Both Section 5.2 and Section 5.3 start with a discussion of the analysis of clinical data, and then followed by a discussion of the analysis of simulated data.

Section 5.4 is a discussion of issues in the analysis of agreement and reliability found in this study. The limitations of this study will be discussed in Section 5.5. Finally, Section 5.6 is the summary of the findings in this study.

5.2 Analysis of Agreement

5.2.1 Analysis of Clinical Data

5.2.1.1 Comparison of prediction of agreement

All instruments used in this study were validated by their manufacturer and have been used in a clinical setting. Therefore, all instruments should be in agreement with their standard. However, as presented in the Chapter 4 (Section 4.3.1), only the agreement model and ICC_A show agreement for all of the instruments. The comparing slopes and y-intercepts analysis only shows agreement for instruments measuring SBP, and not for those measuring other variables (i.e. glucose level, diastolic blood pressure, weight, and peak expiratory flow rate). The Bland-Altman LoA shows disagreement for all instruments.

The comparing slopes and y-intercepts analysis involves testing the slope and intercept of the line ($y=\alpha+\beta x$) with the line of agreement ($y=x$). The agreement of the data set only concluded when the slopes and y-intercepts are equal (i.e. when the two lines coincide). As discussed earlier in the Chapter 2 and Chapter 3, the method used to fit the line was an Ordinary Least Square (OLS) method. The ordinary least squares method is known to be very sensitive to outliers (Bashiria & Moslemia, 2011; Bilić-Zulle, 2011; Ugrinowitsch, Fellingham, & Ricard, 2004). In a data set with outliers, the task of outlier detection is challenging, and masking may occur. Furthermore, for two lines ($y=\alpha+\beta x$ and $y=x$) to be equal, the intercepts of the two parallel lines should be equal. Therefore, even with very small differences in the intercepts of the two lines, the differences of the two lines will be significant. This makes the comparing slopes and y-intercepts analysis is a very sensitive method for the detection of small disagreements, regardless of whether the differences are clinically significant or not.

The Bland-Altman analysis failed to show agreement for all of the instruments. Hopkins (Hopkins, 2004) suggested that the Bland-Altman method tends to overestimate bias. This could explain why the Bland-Altman method concluded that there was no agreement for all of the instruments tested in this study, and why bias predicted in the analysis could be overestimated. As a result, there is still the possibility of an agreement when the Bland-Altman analysis concluded with a disagreement of instruments. Conversely, when the Bland-Altman analysis resulted in the agreement of instruments, it is very likely that the instruments are truly in agreement.

Another possible explanation for this finding is that the distribution of the differences for all datasets failed the normality test. In the analysis, the D'Agostino and Pearson omnibus normality tests were used to test for the normality. However, the shapes of the distributions were approximately bell shaped for all of the variables. Bland and Altman suggested that the distribution should be approximately normal, which can be checked using histograms and a normal quantile plot of the differences (Bland & Altman, 1986; Bland & Altman, 2012). They did not specify whether variables have to pass any specific normality tests, and they also stated that this normality assumption does not have to be met closely, and it is unlikely to be a problem if the variability of the differences is constant (Bland & Altman, 2012).

The distribution of the differences and the LoA can be clearly seen in the Bland-Altman plot. However, the Bland-Altman plot alone provides only limited information (M. W. Smith, Ma, & Stafford, 2010). A Bland-Altman plot will reveal outliers, but without their associated frequencies it is difficult to interpret (M. W. Smith et al., 2010). Therefore, a bar chart or a histogram of the differences is suggested to complement the Bland-Altman plot (Altman & Bland, 1983; M. W. Smith et al., 2010). This was actually proposed by Altman and Bland in one of their early articles (Altman & Bland, 1983), but has rarely been practiced.

To illustrate the application of the histogram, the clinical data for glucose is used. From the Bland-Altman analysis, the LoA ranges from -0.62mmol/l to 1.24mmol/l. This means that the glucometers tend to overestimate the laboratory values by 1.24mmol/l and underestimate the laboratory values by 0.62mmol/l. The clinically significant difference was set as 0.8mmol/l at the beginning of this study (Essack et al., 2009). The upper limit of agreement suggests that the differences of readings between the glucometers and the laboratory values can exceed the acceptable differences (i.e. >0.8mmol/l). This suggests that there is no agreement between the glucometer and the laboratory value. Figure 5.1 is the Bland-Altman Plot for the analysis, and Figure 5.2 is the histogram of the differences. The histogram suggests that about 92% (276 of 300) of the differences were within the acceptable value (between ±0.8mmol/l). If the glucometer is required to give accurate measurements (in agreement with the laboratory values) for 95% of the time, then a 92% accuracy is not acceptable. Therefore, there is no agreement between the glucometer and the laboratory value. This supports the conclusion made from the interpretation of LoA earlier. However, if the threshold is lowered (i.e. if the glucometer is required to give accurate measurements for 90% of the time), then there is an agreement between the glucometer and the laboratory value.

Although all of the instruments used were validated by the manufacturer, no information was given regarding how the process of validation was performed and what acceptable range of differences was set during the validation study. The conclusion on agreement is subjective and unique for different situations. The acceptable range of differences can also be different for different situation. The clinically significant differences set in this study might not match the value set during the validation process by the manufacturers. As an example, the acceptable range for weight set in this study was 0.5kg; however the value set during the validation study might be different.

This could explain the conclusion on the disagreement of all instruments by the Bland-Altman analysis found in this study. This finding is not unique, as a study in South Africa also found that some of the blood glucose meters that have been used in clinical practice gave results that were not comparable to laboratory values (Essack et al., 2009). Details of that study have been discussed in the Chapter 1.

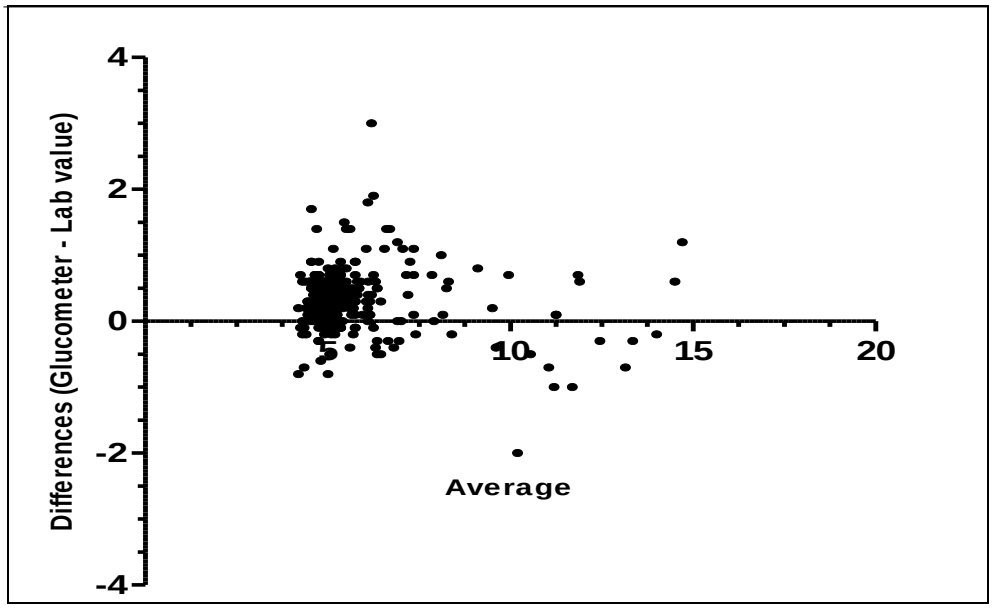


Figure 5.1: The Bland-Altman Plot for Glucose

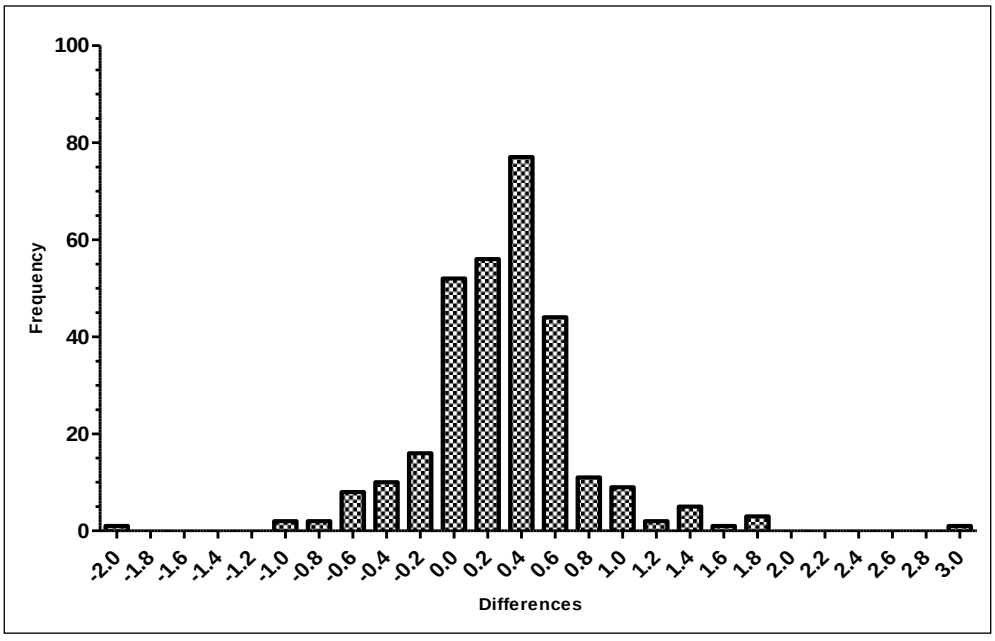


Figure 5.2: Histogram of the differences for Glucose

5.2.1.2 Consistency of prediction

All of the statistical methods provide a consistent prediction of agreement except for the comparing slopes and y-intercepts analysis. The comparing slopes and y-intercepts analysis only provides a consistent prediction of agreement for weight and PEFR. As discussed in the previous section (Section 5.2.1.2), the slope and intercept fitted using the ordinary least squares method is very sensitive to outliers. This may have caused the inconsistency of prediction for the comparing slopes and y-intercepts analysis. Since the comparing slopes and y-intercepts analysis is sensitive, when the results suggest that there is no agreement between the instruments, there is still a possibility that the instruments are truly in agreement.

5.2.2 Analysis of Simulated data

5.2.2.1 Constant systematic error

All four methods correctly predict the disagreement in the simulated data set of all variables. The agreement model and the Bland-Altman LoA correctly predict the magnitude and direction of simulated error in the entire data set. The y-intercept (in the comparing slopes and y-intercepts analysis) also seems to reflect the error in the entire data set. The actual value of ICC_A did not provide a good prediction of agreement, but the confidence intervals (CIs) reflect the disagreement in the dataset. The information from ICC_A analysis also fails to provide information on the direction of error (positive or negative error) in the dataset. The patterns of prediction for all four methods were similar for both positive and negative error.

The nature of constant bias in the simulated dataset suggests that the distribution of biases is not normal. This means that the normality assumption for the Bland-Altman analysis has been violated. Despite this violation, the Bland-Altman methods correctly

predict the simulated bias. This issue has been commented on by Bland and Altman in their recent article (Bland & Altman, 2012). The normality assumption in the Bland-Altman analysis does not have to be closely met if the variability of differences is constant (Bland & Altman, 2012). In the case of constant bias, it is obvious that the variability of the differences is constant.

5.2.2.2 Inconsistent error

The Bland-Altman LoA correctly predicts the disagreements of datasets for all variables. The other methods only detect disagreement for some of the variables. Although the Bland-Altman LoA successfully detects the disagreement, the predicted biases were overestimated. None of the methods were able to quantify the simulated bias correctly.

The overestimations of bias by the Bland-Altman method were seen in all of the variables and the entire simulated dataset. Results shown in the Chapter 4 suggest that there are specific patterns in the prediction of bias by Bland-Altman analysis. The predictions of bias were influenced by different mixtures of positive and negative bias in the dataset. Table 5.1 shows the pattern of prediction for Bland-Altman analysis.

The analysis of datasets with simulated inconsistent bias did not satisfy the normality assumption for the Bland-Altman analysis. This is due to the mixture of fixed positive and negative errors in the dataset. However, the analysis shows how the mixture of these biases influences the prediction of the Bland-Altman LoA.

Table 5.1: Pattern of prediction by the Bland-Altman method

	1/3 positive error and 2/3 negative error	**1/2 positive error and 1/2 negative error**	**2/3 positive error and 1/3 negative error**
Glucose Error = ±0.8mmol	Bias = -0.27mmol/l LoA = -1.75mmol/l to 1.21mmol/l Range of predicted bias = 2.96mmol/l	Bias = 0mmol/l LoA= -1.57 to 1.57mmol/l Range of predicted bias = 3.14mmol/l	Bias = 0.27mmol/l LoA = -1.21 to 1.75mmol/l Range of predicted bias = 2.96mmol/l
SBP Error = ±10mmHg	Bias = -3mmHg LoA = -22mmHg to 15mmHg Range of predicted bias = 37mmHg	Bias = 0mmHg LoA = -20mmHg to 20mmHg Range of predicted bias = 40mmHg	Bias = 3mmHg LoA = -15mmHg to 22mmHg Range of predicted bias = 37mmHg
DBP Error = ±10mmHg	Bias = -3mmHg LoA = -22mmHg to 15mmHg Range of predicted bias = 37mmHg	Bias = 0mmHg LoA = -20mmHg to 20mmHg Range of predicted bias = 40mmHg	Bias = 3mmHg LoA = -15mmHg to 22mmHg Range of predicted bias = 37mmHg
Weight Error = ±0.5kg	Bias = -0.17kg LoA = -1.09kg to 0.76kg Range of predicted bias = 1.85kg	Bias = 0kg LoA = -0.98kg to 0.98kg Range of predicted bias = 1.96kg	Bias = 0.17kg LoA = -0.76kg to 1.09kg Range of predicted bias = 1.85kg
PEFR Error = ±40 lmin	Bias = -13l/min LoA = -87l/min to 61l/min Range of predicted bias = 148 l/min	Bias = 0l/min LoA = -79l/min to 79l/min Range of predicted bias = 158 l/min	Bias = -3l/min LoA = -61l/min to 87l/min Range of predicted bias = 148 l/min

From Table 5.1, it can be seen that when the proportion of negative error is more than the positive error in a single data set, the LoAs seem to show more underestimation of the actual value in comparison with an overestimation of the actual value. The pattern is similar when more positive errors are present in the dataset, where the LoA will show more overestimation of the actual value. When the proportion of positive and negative error is equal in the dataset, the LoAs seem to be symmetrical, and the estimated range of bias seems to be doubled in comparison with the actual error. The mean biases predicted by the Bland-Altman method were all zero. This is because the poor agreements between the two datasets were hidden in the distribution of differences. This

suggests that the Bland-Altman method is influenced by the distribution of error, and that the mean bias estimated in the Bland-Altman analysis cannot be used to make the conclusion on agreement.

This finding shows the importance of satisfying the normality assumption in the Bland-Altman analysis. However, there should be an exception on the normality assumption when the bias is constant, or when the variability of differences is constant (Bland & Altman, 2012). The Bland-Altman method has been shown to be good in predicting constant bias (as shown in Section 5.2.2.1). Instruments which produce inconsistent bias are not precise or not reliable. The issue of inconsistent bias produced by an instrument should not be a cause for concern if the instrument is reliable. Therefore, it is important to ensure the reliability of certain instruments before testing its accuracy.

5.2.2.3 Proportion of bias

The comparing slopes and y-intercepts analysis and the Bland-Altman analysis successfully predict the disagreement in the entire dataset for all of the variables. However, the comparing slopes and y-intercepts analysis did not provide information on the direction or quantification of error for all of the variables. Furthermore, the comparing slopes and y-intercepts analysis is known to be highly sensitive to the distribution of error, especially outliers, and, as discussed in the Section 5.2.1, the prediction of bias in the comparing slopes and y-intercepts analysis is not consistent. The prediction of bias in the Bland-Altman analysis was also overestimated. The agreement model and the ICC_A only detected disagreement for some of the variables. Supposedly, when only part of the dataset has bias, the situation is actually similar to the dataset with inconsistent bias. Therefore, by ensuring the reliability of an instrument, the problem of detecting inconsistent bias can be avoided.

5.2.2.4 Sample size

The estimation of sample size for the regression analysis can be calculated by the formula suggested by Cohen (J. Cohen, 1977), as discussed in Chapter 2. Therefore the effect of sample demonstrated in Section 4.3.2.4 is mostly relevant to the Bland-Altman analysis and the ICC_A.

The predictions of outcome for all four methods (comparison of slopes and y-itercepts, agreement model, Bland-Altman LoA, and ICC_A) were influenced by the sample size. The standard errors of the prediction for all methods stabilise when the sample size is greater than 100. The pattern of prediction became even more constant when the sample was greater than 200. The standard error is an index of the variability of the means that would be expected if the study were repeated a large number of times (Altman & Bland, 2005; Streiner, 1996). Since the standard error becoming very constant when the sample size is greater than 200, there is little point in having a sample size greater than 200 in agreement studies when using the Bland-Altman method and ICC_A. No pattern or variation was detected in the effect of sample size for a different range of variables.

Doros and Lew (2010) have suggested a sample size estimation for ICC based on the expected width of the confidence interval. They estimated that for a 95% CI, a sample size of more than 50 is required to detect an ICC $\geq$ 0.6, and a larger sample size is required to detect a much smaller ICC. However, this estimation was based on three repeated measurements. From the findings in the Chapter 4, a sample size of 100 is reasonably safe for analysis using ICC_A (based on two repeated measurements).

The findings in this study also support the sample size suggested for Bland-Altman analysis (Bland, 2004). By estimating the confidence intervals of the limits of agreement, Bland (Bland, 2004) recommended a sample size of 100 for the analysis,

but also added that a sample size of 200 is even better, but this depends on what accuracy is required by the researcher. Despite the recommendation, the importance of sample size in the Bland-Altman analysis seems to have been neglected. In the systematic review earlier in the Chapter 2, out of the 178 agreement studies that have used the Bland-Altman method, 60% have sample sizes of less than 100, and about 50% of the studies have sample sizes less than 50, with the most popular sample size being 30. Figure 4.26 (Section 4.3.2.4) demonstrates the impact of a low sample size on the prediction of LoA. When the sample size is 30, the prediction of lower limit of agreement for DBP is between -23mmHg and -9mmHg, and the prediction of the upper limit of agreement is between 11mmHg and 28mmHg.

Since the Bland-Altman method is the most commonly applied method to assess agreement, the issue on appropriate sample size in the Bland-Altman analysis, and the effect of low sample size on the prediction of bias and limits of agreement needs to be highlighted. Otherwise, agreement studies conducted with low sample sizes will produce an inaccurate prediction of error. This might affect the quality of medical instruments used in clinical practice, and could potentially affect the quality of care given to the patient.

5.2.3 Extended analysis of the Bland-Altman method

5.2.3.1 Proportional Bias

One of the critiques of the Bland-Altman analysis is the existence of proportional bias. Proportional bias is present when the difference in values resulting from two methods or instruments increases or decreases in proportion to the average values of the two measurements. As discussed in detail by Hopkins (2004), the Bland-Altman plot causes an artefactual bias. For a standard instrument, A, and a comparison instrument, B, the differences of the measurement A-B tends to be positive for the larger average values of

(A+B)/2 and negative for the smaller values. This is because when A-B>0 (i.e. A>B) the average reading will be greater than when A-B<0 (i.e. A<B) (Hopkins, 2004). This means that the Bland-Altman plot will indicate that bias is present even when there is no bias. Therefore, part of the bias predicted by the Bland-Altman analysis will be artefactual, and the researcher will not be able to differentiate whether the bias is an artefact or real (Hopkins, 2004).

To overcome the problem of proportional bias in the Bland-Altman analysis, it is recommended that a linear regression line (differences in readings against the mean of readings) should be fitted to the Bland-Altman plot (Ludbrook, 2010). If the slopes of the line do not significantly differ from zero then the proportional bias is absent (Ludbrook, 2010). The analysis in Section 4.3.3 shows that there is an overestimation in the prediction of bias in the Bland-Altman analysis even when the proportional bias had been excluded. The overestimation of bias increases when the range of actual bias increases. The application of confidence intervals for the LoA, makes the problem with overestimation of bias even worse. Similar patterns were found for the results of all three variables which meant that this was not an isolated issue.

Excluding the proportional bias using the regression line analysis of the Bland-Altman plot does not remove the possibility of artefactual bias in the prediction of bias. The regression line analysis of the Bland-Altman plot cannot be used to exclude the proportional bias in the analysis. The issue of overestimation bias or artefactual bias in the Bland-Altman analysis should be highlighted to researchers.

5.2.3.2 Confidence Intervals for the Limits of Agreement

The tendency of the limits of agreement to overestimate bias also raised another issue: whether the application of confidence intervals for the limits of agreement in the Bland-Altman analysis is really necessary. The limits of agreement are sample estimates, thus

Bland and Altman (Bland & Altman, 1987) suggested that 95% confidence intervals of the upper and lower limits of agreement should be calculated. However, these confidence intervals are hardly reported. Out of the 178 papers that have used the Bland-Altman method reviewed earlier in Chapter 2, only one paper reported the 95% confidence intervals of limits of agreement. Bland and Altman are also aware of this problem and regret that these confidence intervals are seldom reported (Bland & Altman, 2003). In their recent article (Bland & Altman, 2012), they have highlighted this issue again.

In 2007, based on simulated data, Hamilton and Stamey estimated the probability that the limits of agreement alone will actually contain 95% of the population from which the fictitious differences are drawn (Hamilton & Stamey, 2007). They found that the limits of agreement contain 95% percent of the distribution less than 65% percent of the time, even for a sample size of 200 (Hamilton & Stamey, 2007). They concluded that reporting the LoA without corresponding them to the confidence intervals is rather like reporting a sample mean without a confidence interval (Hamilton & Stamey, 2007). However, it is obvious that the confidence intervals will always contain values that are more than the estimated parameters.

Although the importance of confidence intervals of the LoA has been highlighted, the conclusion on the agreement of instruments has always been interpreted from the LoA. In their paper proposing the LoA (Bland & Altman, 1986), Bland and Altman also relied on the LoA in the interpretation of agreement themselves. So, what is the significance of reporting the confidence intervals, when the interpretation of agreement is still based on the limits of agreement? The LoA itself is actually an interval of an estimate parameter. According to the formula for calculating the LoA (Bland & Altman, 1987):

Limits of Agreement = mean difference ± 1.96 x (standard deviation)

where the value of 1.96 is the z-score based on the estimation of the standard normal probability distribution (two-tailed probabilities of 95% of differences will lie between these limits). Therefore, the LoA is an interval for the estimated bias (differences between two measurements) and 95% of differences should lie within these limits. It gives an estimated range of differences or biases which is likely to include an unknown population parameter.

Since the interpretation of the Bland-Altman method is based on the LoA, the importance of reporting confidence intervals for the LoA should be revised. Furthermore the LoA itself is an interval and already gives a range of possible biases produced by any tested instruments. Therefore, further study to test the importance of confidence interval for the LoA is needed to clarify this issue.

5.3 Reliability Analysis

5.3.1 Clinical data

5.3.1.1 Comparison of prediction

All of the instruments used in this study were validated by their manufacturer and have been used in a clinical setting. Therefore, all instruments should be reliable, and the ICC_A and ICC_C predict that all instruments were reliable. In contrast, the Bland-Altman method suggests that all of the instruments were not reliable, except for the infrared thermometer (to measure body temperature).

The ICC_A was initially constructed to be used for the analysis of agreement. However, this method also provides a similar conclusion on the prediction of reliability with the ICC for consistency (ICC_C), although the actual value of the ICC_A tends to be lower than the ICC_C. The extra parameter in the denominator for the ICC_A formula in comparison with the ICC_C formula explains the tendency of the ICC_A giving a lower

estimation of the ICC parameter. Based on the estimation from analysis of variance (ANOVA), the formula for ICC (A, 1) or ICC (2, 1), which is used for agreement, is given as follows (Weir, 2005):

$$\mathrm{ICC} = \frac{\mathrm{MS}_S - \mathrm{MS}_E}{(\mathrm{MS}_S + (k-1)\mathrm{MS}_E) + \left(k\frac{\mathrm{MS}_T - MS_E}{n}\right)}$$

Whereas the formula for ICC (C,1) for consistency or reliability is given as follows (Weir, 2005):

$$\mathrm{ICC} = \frac{\mathrm{MS}_S - \mathrm{MS}_E}{(\mathrm{MS}_S + (k-1)\mathrm{MS}_E)}$$

The Bland-Altman analysis failed to predict the good reliability of all instruments, except for the infrared thermometer. However, it was shown earlier (in Section 5.2) that the Bland-Altman method has a tendency to overestimate bias. The use of LoA to evaluate reliability has been criticised in the past (Hopkins, 2000), due to the overestimation of bias, and cannot be applied to the simplest situation of only one trial as discussed in Chapter 2. Furthermore, the Bland-Altman analysis was designed for the analysis of two repeated readings. This breaches the concept of reliability, which allows unlimited repeated numbers of observations per subject (Fay, 2005).

Bland and Altman (1999) have suggested a method for the application of multiple measurements for the analysis. They also proposed calculating the mean of the replicated measurements by each instrument, for each subject (Bland & Altman, 2007). These pairs of means could be used to compare the two instruments using the limits of agreement (Bland & Altman, 2007). However, this was only suitable for the analysis of agreement. In reliability analysis, the reading of each repeated measurement is important so this makes the Bland-Altman analysis unsuitable for the analysis of reliability.

5.3.1.2 Consistency of prediction

The results in Section 4.4.1.2 show that all three methods provide a consistent prediction of reliability for all ten sets of data for all of the variables. Both ICC_A and ICC_C predict the good reliability of all of the tested instruments. The value of ICC_A was also found to be lower than the value of ICC_C for a similar set of data. As explained in the previous section (Section 5.3.1.1), this was due to the extra parameters in the denominator of the ICC_A formula.

5.3.1.3 Number of measurements

Results in section 4.4.1.3 show that both ICC_C and ICC_A provide similar predictions of reliability with two and three repeated measurements. All of the tested medical instruments were reliable, so this is probably the reason why both predictions using ICC_C and ICC_A were similar. There is no error or imprecision of data to be detected by these methods. However the range of confidence intervals for prediction with three repeated measurements is smaller than the prediction with two repeated measurements. This suggests that the predictions of both ICCs based on three repeated measurements are more precise than two repeated measurements. The pattern is similar for all of the variables.

5.3.2 Simulated data: comparison of prediction

The simulated data were designed so that the five repeated measurements were not consistent (i.e. not reliable). Results of the analysis in Section 4.4.2 show that ICC_C failed to detect the poor reliability for all the simulated data. In contrast, the ICC_A predicts the poor reliability of all of the analysis except in the analysis with two

repeated measurements. The pattern of findings is similar for all of the tested variables. Although the ICC_C was designed to test for consistency (i.e. reliability), the results suggest that the ICC_A is actually better for the prediction of the reliability of instruments measuring a continuous outcome. As explained earlier, the main differences between the formulae for calculating the ICC_A and ICC_C are in the denominator. A simple data of instrument measuring variable 'V' from Table 5.2 will help to explain this finding.

Table 5.2: Data to demonstrate the differences between ICC_A and ICC_C

Subject	1st reading	2nd reading	3rd reading
A	1	2	3
B	2	3	4
C	3	4	5
D	4	5	6
E	5	6	7

From Table 5.2 it is obvious that the repeated readings for those subjects were not the same. This means that the instruments measuring the variable 'V' is not reliable. The analysis using both SPSS and MedCalc software will produce the same answer of ICC_C = 1.00, and ICC_A = 0.7143 (0.0599 to 0.9638). These ICC values can be calculated manually. The ANOVA table for the data from Table 5.2 is shown in Table 5.3.

Table 5.3: ANOVA table for analysis of variable 'V"

	Degree of freedom (df)	Sum of Squares (SS)	Mean Square (MS)
Between Subject (S)	4	30	7.5
Within Subject	10	10	1
• Trial (between items, T)	2	10	5
• Error (residual, E)	8	0*	0*
Total	14	40	2.857

**very small approximately equal to zero*

According to the formula, to calculate the ICC_C (ICC for consistency or reliability):

$$ICC_C = \frac{7.5 - 0}{(7.5 + (3 - 1)0)} = \frac{7.5}{7.5} = 1.0$$

To calculate the ICC_A (ICC for agreement):

$$ICC_A = \frac{7.5 - 0}{(7.5 + (3 - 1)0) + \left(3\frac{5 - 0}{5}\right)} = \frac{7.5}{7.5 + 3} = 0.714$$

The formula for ICC_A takes into account the value of MS_T (mean square of trials). This means that the ICC_A formula considers the variability due to differences between repeated measurements. The formula for ICC_C only takes into account the variability due to error, and the between-subject variability. In other words, the ICC for consistency simply compares the consistency between trials, whereas the ICC for absolute agreement compares both the consistency between trials and the agreement between ratings (Weir, 2005).

In clinical practice, when the value of certain continuous variables is measured (such as blood glucose level or haemoglobin level), it is expected that repeated readings of the same patient would give exactly the same value (or very small differences in the repeated readings). This suggests that the differences between repeated readings (agreement between ratings) should be considered when testing the consistency or reliability of an instrument measuring continuous variables.

Therefore, the ICC_A should be the choice of analysis for testing the reliability of instruments measuring continuous variables. The term 'ICC for absolute agreement' and 'ICC for consistency' should not confuse researchers and influence their decisions when choosing which type of ICC should be used in testing the reliability of instrument. However, this is only applied to the instruments measuring continuous variables, and

this is not the case for testing instruments measuring other outcomes (such as ordinal or nominal variables).

Although the ICC_A actually provides a more accurate prediction on the reliability of instruments, the interpretation of reliability should be based on the confidence intervals (not the single ICC parameters). As shown in Section 4.4.2, all of the ICC_A values suggest that the measurements were reliable. However, the confidence intervals for all of the ICC_A values (except predictions with two repeated measurements) were able to detect the imprecision in the datasets. The ICC is just an estimated reliability parameter, so the confidence intervals will give a range of possible true values.

The importance of confidence intervals for the ICC should be highlighted to all medical researchers because the confidence intervals for ICC are sometimes not reported, and the conclusion of reliability is only made based on a single ICC parameter. Out of 24 reliability studies that have used the ICC found in the systematic review earlier in Chapter 2, 46% do not report the CIs, and were concluded based on a single ICC value.

5.3.3 Extended analysis of the ICCs

5.3.3.1 Clinical data: Comparison of two and three repeated readings

Results in Section 4.4.1.3 suggest that the conclusions on reliability were similar for both three and two repeated measurements for both ICCs. However, the confidence intervals for ICCs based on three repeated measurements were smaller compared to the confidence intervals of ICCs with two repeated measurements. No significant differences were detected between the predictions of ICCs using two or three repeated measurements. Both p-values for ICC_A and ICC_C were greater than 0.05. As explained

in Section 5.3.1.3, this is probably due to no error in the dataset to be detected by those methods.

5.3.3.2 Simulated data of poor reliability instruments

The analysis in Section 4.4.2 also shows that there are differences in the prediction of reliability using a different number of repeated measurements (two, three, four, and five repeated measurements) for ICC_A. Only the prediction of reliability based on three or more repeated measurements accurately predicted the poor reliability of the dataset. The extended analysis in section 4.4.3.2 shows that the differences in the prediction of reliability based on two and three repeated measurements were significantly different for both ICC_A and ICC_C. Therefore, to test for the reliability of instruments, it is important to have at least three repeated measurements for each subject.

5.4 Issues in method comparison studies

5.4.1 Agreement or Reliability?

Agreement signifies the accuracy of certain instruments, whereas reliability indicates precision. Preferably, these parameters should be assessed together in a validation study. However, as found in the systematic review in Chapter 2, it is not commonly followed in practice, especially with respect to agreement studies. Only 30% of the agreement studies assessed reliability, as compared to 75% of the reliability studies that also measured agreement at the same time.

Researchers tend to only focus on one aspect of quality in validating instruments, although there is a possibility of agreement and reliability studies being conducted separately for the same instrument. In fact, as discussed in Chapter 1, both agreement and reliability are important in determining the quality of the instrument.

Since no method is able to correctly predict inconsistent error, this problem can be avoided by ensuring the instrument only produces constant bias. This only happens if the instrument is giving precise measurements (i.e. reliable).

Therefore, it is important for researchers to test for the reliability of the instrument before testing its accuracy or agreement with the standard, or at least report whether the instrument has been shown to be reliable. It is impossible to assess the agreement of imprecise instruments, and it useless to have a precise instrument that gives inaccurate measurements.

5.4.2 Single or Multiple methods?

According to both systematic reviews conducted in Chapter 2, most of the reliability studies (87%) relied on a single statistical method to assess reliability, in contrast with agreement studies, where most of the studies (65%) used a combination of statistical methods. The findings in this study suggest that all of the statistical methods that have been used to test for agreement have their own strengths and weaknesses. No single method is powerful enough to detect the accuracy of instruments, except for when the error produced is constant. The use of multiple methods has the advantage of compensating for the limitations of any single method (Luiz & Szklo, 2005). However, the application of multiple inappropriate statistical methods, for example the use of both correlation coefficients and significance tests of the difference between means, should be avoided (i.e. should not be used).

Bruton et al. (2000) suggested that no single reliability estimate should be used for reliability studies, and that a combination of methods was more likely to give more information on the reliability of an instrument. However, the use of a single method is more popular in the analysis of reliability. The use of ICC to test for reliability was found to be the most popular single method used to test reliability. Despite the

popularity of this method, it is important for researchers to be aware of different types of ICCs and the different formulae used to compute ICC.

Although ICC for consistency (ICC_C) was associated with the analysis of reliability (to test for consistency), this study shows that the ICC_A is actually the method to be used for analysing the reliability of instruments with continuous outcomes. The ICC_A was thought to be used for the analysis of agreement, but the inability of this method to quantify bias makes it less suitable to use for the analysis of agreement.

5.4.3 Application of Inappropriate Statistical Methods

The systematic reviews in Chapter 2 show that 19% of the reliability studies and 10% of the agreement studies used inappropriate methods. This means that there is a possibility that some medical instruments or equipment currently used were validated using inappropriate methods with consequently erroneous conclusions being drawn from these methods. Therefore, this equipment may not truly be as precise or accurate as believed.

The question of which method is the best or most appropriate is also difficult to answer because there is no single perfect method. Different methods have different strengths and weaknesses. Therefore, the issue now is which combination of methods is the most appropriate to test for agreement and reliability of instruments. In fact, both agreement and reliability should be assessed together in a method comparison study. A flow of analysis with a combination of methods will be proposed in the next chapter (Chapter 6).

5.4.4 Need for guidelines

Both systematic reviews in Chapter 2 suggest that there is a gap in the knowledge among medical researchers in this area. There were inappropriate applications of statistical methods in the analysis of agreement and reliability, and the reliability of instruments was not measured or not reported in most of the agreement studies.

Recently, the guidelines for reporting reliability and agreement studies (GRRAS) have been proposed (Kottner et al., 2011). These guidelines found that the reporting of method comparison studies (both agreement and reliability studies) were incomplete and inadequate. Information about sample selection, study design and statistical analysis were often incomplete (Kottner et al., 2011).

Even a recent article (Hanbazaza & Mansoor, 2012) relied on the use of inappropriate analyses to test for agreement. Thus, guidelines on how to perform the analysis in method comparison studies are really needed. Furthermore it is also important to educate medical researchers and clinicians on the concept and analysis in method comparison studies.

5.5 Limitation of study

There are several limitations at different stages of this study:

1. This study only looked at the most commonly used statistical methods in medicine, as found in the systematic review. There are other methods that have been used in method comparison studies, especially in the analysis of agreement, such as the Passing and Bablok regression method (Bilić-Zulle, 2011; Passing & Bablok, 1983), and a graphical approach suggested by Luiz et. al (Luiz, Costa, Kale, & Werneck, 2003).
2. The results of both systematic reviews also have limited generalisation due to selection bias. These reviews were limited to five electronic databases

(Medline, Ovid, PubMed, Science Direct and Scopus) and were also limited to articles published only in English. The searches were only performed using online databases, and, as such, unpublished articles were not considered. However, these databases have a very wide coverage of published medical journals including high quality and high impact journals. A broad search term was used for each systematic review, in order to capture the largest possible number of publications on the topic. Two independent reviewers were also used during the selection of articles and data extraction in order to reduce bias.

3. The analysis of this study also limited to the variables collected in the study. However, a wide range of variables were collected: five variables (blood glucose level, SBP, DBP, body weight and PEFR) for agreement analysis and six variables (SBP, DBP, heart rate, body weight, body temperature and PEFR) for reliability analysis.
4. The ranges of variables collected were not wide enough to cover all ranges of extreme values (i.e. very low and very high values). The interpretation of results of the analysis especially for regression analysis cannot be cannot extrapolated, thus limited to available data. However, effort has been put into obtaining a wide range of values for each variable by collecting data from multiple centres and different study populations. The data collected covers a good range of normal values and some abnormal values, except for the body temperature, which has a very limited range of values.
5. Most of the participants were Malay, however the race of the participants was not an important issue in this study, as only the variability and range of data collected for each variable is important.

6. Although the issue on inter-observer error has been avoided by ensuring that only one person (i.e. the researcher) takes all of the measurements during data collection, there was also an issue regarding intra-observer error. However, this error was reduced by ensuring that the researcher followed the standard procedure set for the measurement of each variable (see chapter 3 for the detail of measurement procedures).

5.6 Summary

5.6.1 Agreement Analysis

This chapter begins with the discussion of results for the analysis of agreement in Section 5.2. All four methods provide different predictions of bias and different conclusions on agreement for a similar set of clinical data. The agreement model, Bland-Altman LoA and ICC_A provide a very consistent prediction of agreement, but not for the comparing slopes and y-intercepts analysis. This was most likely due to the sensitivity of the ordinary least squares method to outliers.

The analysis of simulated data suggests that all methods are appropriate for predicting agreement when the bias is constant. However, the ICC_A is unable to quantify the magnitude and direction of bias. This makes the ICC_A the least useful method for testing the agreement of instruments. Although the agreement model provides an estimation of minimum and maximum bias of an instrument, their results cannot be extrapolated and are limited to the range of data included in the analysis.

Results for the prediction of inconsistent bias were discussed in Section 5.2.2.2. Only the Bland-Altman method correctly predicts the disagreement of the simulated dataset with inconsistent bias for all variables. However, the biases predicted were overestimated. This suggests that no method is good for predicting inconsistent bias,

because all methods are greatly influenced by the direction or distribution of error. In addition, the mean bias produced by the Bland-Altman method should not be used to assess agreement. The interpretation of Bland-Altman analysis should be based on the limits of agreement (upper and lower limits). Section 5.2.2.2 also discussed the importance of satisfying the normality assumption in the Bland-Altman analysis, except when the variability of differences is constant.

Section 5.2.2.3 discussed the findings for analysis with different proportions of bias in the dataset. The situation of where only part or proportion of dataset has bias, suggests that the instrument is producing inconsistent error. Since there is no way of identifying the inconsistent error of an instrument, it is important to test the reliability of an instrument first before checking its accuracy. It is impossible to test the accuracy of an unreliable instrument.

The issue of sample size was discussed in Section 5.2.3. The predictions of outcomes for all methods were influenced by sample size. The estimation of sample size for the comparing slopes and y-intercepts analysis can be calculated by a formula. However, the estimation of sample size for the Bland-Altman method and the ICC_A were based on precision. The standard errors of the prediction for all methods stabilise when the sample size is greater than 100. The pattern of prediction becomes even more constant when the sample is greater than 200. Therefore, a minimum sample size of 100 is required for the Bland-Altman analysis and the ICC_A, but a sample size of more than 200 is considered to be a waste of resources. This section also highlighted that the issue of the appropriate sample size for the Bland-Altman analysis was neglected by researchers, and shows the impact of low sample sizes in the prediction of bias using Bland-Altman analysis.

Section 5.2.3 discussed the extended analysis of the Bland-Altman method. Two main issues discussed in this section were the proportional bias in the Bland-Altman

analysis and the confidence intervals for the limits of agreement. It has been shown that the Bland-Altman analysis tends to overestimate bias, most likely due to the proportional bias in the analysis. The regression analysis of the Bland-Altman Plot was proposed to exclude the proportional bias in the Bland-Altman analysis (Hopkins, 2004). However, the findings in this study showed that the Bland-Altman analysis still overestimates bias even when the proportional bias was excluded using regression analysis. The issue of the overestimation of bias in the Bland-Altman analysis has raised another issue on the significance of reporting the confidence intervals for the LoA. Although the LoAs were thought to be the sample estimates (Bland & Altman, 2012), the formula for LoA suggests that it is an interval estimate, and since the interpretation of agreement is based on the LoA, the importance of reporting the confidence intervals should be revised.

5.6.2 Reliability Analysis

Section 5.3 is a discussion of the results for the analysis of reliability. Statistical methods tested by the reliability analyses were the Bland-Altman Limits of Agreement (LoA), the Intra-class Correlation Coefficient for consistency (ICC_C), and the Intra-class Correlation Coefficient for agreement (ICC_A). Section 5.3.1 discusses the analysis of clinical data. Both ICC_C and ICC_A predict the reliability of all instruments. The fact that the Bland-Altman method failed to provide a good prediction of the reliability of instruments, along with the fact that this method is not suitable for the analysis of data with more than two repeated measurements, means that the Bland-Altman LoA is the least suitable method for the analysis of reliability. Both ICC_C and ICC_A were consistent in the prediction of the reliability of the clinical data, and no differences were found between the predictions of reliability with two or three repeated measurements. This is probably because there was no error to be detected in reliable instruments.

Section 5.3.2 is the discussion of the reliability analysis of simulated data. The analysis shows that the ICC_A is actually better than the ICC_C for the prediction of reliability. The formula for ICC_A takes into account the variability due to differences between repeated readings (mean square of trials, MS_T). The ICC_A should be the choice of analysis for testing the reliability of instruments measuring continuous outcome. In addition, the interpretation of ICC should be based on the confidence intervals, not on the ICC value. Section 5.3.3 discussed the finding from the extended analysis of reliability. The analysis of simulated data shows that there were significant differences between the predictions of reliability based on two and three repeated measurements. This suggests that it is important to have at least three repeated measurements (for each subject) when testing the reliability of any instruments.

5.6.3 Other Issues and Limitation

Section 5.4 is a discussion of issues in the analysis of agreement and reliability found in this study. Four main issues were discussed in this section. First was the issue of testing both the agreement and reliability of instruments in a method comparison study. This section highlighted the importance of testing the reliability of instruments before testing the agreement. The next issue was the application of multiple methods in the analysis of agreement and reliability. The application of multiple methods was mainly found in the agreement studies, and this was mainly due to the limitation of each statistical method. However, in the reliability studies, the analysis seems to be dominated by the ICC. The third issue discussed was the application of inappropriate statistical methods in the analysis. Since there is no single method that is perfect, and both the reliability and agreement of instruments should be assessed together, it is important to identify the most appropriate combination of methods to assess the reliability and agreement of instruments in method comparison studies. The last section (Section 5.4.4) highlighted

the needs for recommendations or guidelines in analysing data in method comparison studies.

Finally, this chapter presents the limitations of this study, and discusses how the researcher tried to overcome the problem. These include the limited statistical methods tested in this study, the limitations of systematic reviews, the limitations of the analysis of five variables in the agreement study and six variables in the reliability study, the limited ranges of variables, the fact that the samples of the study were dominated by the Malay race, and also the possibility of intra-observer errors during data collection.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

This study compares the most commonly used statistical methods to test for the agreement and reliability of medical instruments with a continuous outcome. Agreement and reliability are two important parameters in determining the quality of instruments. Agreement signifies the accuracy of certain instruments, and reliability concerns the precision of instrument. These two parameters should be tested in method comparison studies or validation studies.

The issue on how to test for the agreement and reliability of medical instruments with a continuous outcome is studied and discussed in this thesis. This chapter presents conclusions from this thesis and recommendations for the analysis of method comparison studies. Section 6.2 is the conclusion on the analysis of agreement and Section 6.3 is the conclusion on the analysis of reliability. The recommendations are presented in Section 6.4, Section 6.5 presents the major contributions of this study, and suggestions for further work are presented in Section 6.6. Finally, Section 6.7 is the summary of this chapter.

6.2 Conclusion on the analysis of agreement

All of the statistical methods were used for the same purposes (i.e. to test for agreement). In theory they should provide the same conclusions on agreement for a similar dataset. However, as discussed in Chapter 5, different statistical methods provide different conclusions on agreement. Each method has strengths and weaknesses, which are summarised in Table 6.1.

Table 6.1: The strengths and weaknesses of each method

	Strength	**Weakness**
Comparison of slopes and y-intercepts	Good in predicting constant bias - Quantify bias - Identify direction of bias	• Very sensitive with outlier • Inconsistent prediction • Not good with inconsistent bias • Influenced by proportion and direction of bias
Agreement Model	Good in predicting constant bias - Quantify bias - Identify direction of bias	• Not good with inconsistent bias • Influenced by proportion and direction of bias
Bland-Altman LoA	Good in predicting constant bias - Quantify bias - Identify direction of bias	• Not good with inconsistent bias • Influenced by proportion and direction of bias • Tendency to overestimate bias
ICC_A	Good in predicting constant bias	• Does not quantify bias • No clues on the direction of bias • Not good with inconsistent bias • Influenced by proportion and direction of bias • Must rely on the CI (not the actual ICC parameter)

All methods are good at predicting agreement or disagreement when the bias is constant. However, the ICC_A does not provide information on the quantification and direction of bias. This makes this method the least useful for testing the agreement of instruments. The comparing slopes and y-intercepts analysis involves a very sensitive hypothesis test, and means that the prediction from the comparing slopes and y-intercepts analysis is not consistent. No method is good for predicting inconsistent bias, because all of the methods are influenced by the proportion and direction of bias (distribution of bias).

The main problem with the Bland-Altman method is the overestimation of bias in the limits of agreement. The fact that the LoA is an interval explains the tendency of LoA to overestimate bias. However, the problem of proportional bias is also another concern of this method (as discussed in Section 5.2.3.1). The regression analysis of the Bland-Altman Plot will not resolve the problem with the overestimation bias due to proportional bias. The issue of the overestimation of bias also raised the question of the need of reporting confidence intervals for the LoA.

It is also important not to make conclusions based on the mean bias predicted in the Bland-Altman analysis, but it must be based on the LoA. The tendency of this method to overestimate bias suggests that, when the conclusion of agreement is achieved in the analysis (based on the LoA), there is true agreement in the dataset.

The Bland-Altman method gives the most frequent accurate predictions of disagreement (all 45 simulated datasets), and the agreement model gives the most frequent accurate predictions of agreement (all 55 clinical datasets) in Chapter 4 (see Table 6.2). The tendency of the Bland-Altman method to overestimate bias makes this method is most suitable to confirm agreement (i.e when the Bland-Altman analysis

resulted in the agreement of instruments, it is very likely that the instruments are truly in agreement).

The agreement model is less likely to give incorrect prediction of disagreement (see Table 6.2). When the agreement model analysis resulted in the disagreement of instruments, it is very likely that the instruments are truly not in agreement. So this method is most suitable to confirm disagreement.

Table 6.2: Percentage of correct and incorrect prediction

	Correct prediction of agreement (Clinical data)	**Correct prediction of disagreement (Simulated data)**	**Incorrect prediction of agreement (FALSE POSTIVE)**	**Incorrect prediction of disagreement (FALSE NEGATIVE)**
Agreement Model	55/55x100 = 100%	25/45x100 = 56%	40/45x100 = 89%	0/55x100 = 0
Bland-Altman LoA	0/55x100 = 0	45/45x100 = 100%	0/45x100 = 0	55/55x100 = 100%

If the reliability of the instrument is ensured, the problem of inconsistent bias can be avoided. Thus the Bland-Altman method and the agreement model will be able to provide good predictions of bias**.** However, because there is a tendency of the Bland-Altman method to overestimate bias, this method should be used with caution and complemented by other methods such as the agreement model. The use of histograms of differences might provide guidance on how far the estimation of error is from the true value.

6.3 Conclusion on the analysis of reliability

Three methods were compared in the analysis of reliability: the Bland-Altman LoA, the Intra-class Correlation Coefficient for consistency or reliability ($ICC_{C)}$ and the Intra-class Correlation Coefficient for absolute agreement (ICC_A). The Bland-Altman LoA is the least likely to be suitable for the analysis of reliability because this method is not suitable for the analysis of data with more than two repeated measurements. This violates the concept of reliability which allows an unlimited number of repeated measurements per subject (Fay, 2005). The analysis in this study also showed that the Bland-Altman method failed to provide a good prediction on the reliability of instruments (with simulated data).

Both the ICC_C and ICC_A were consistent in the prediction of reliability of the clinical data. Although the ICC_C was thought to be the best method to assess reliability, the ICC_A was found to be better than the ICC_C in the prediction of reliability. The formula for ICC_A takes into account the variability due to differences between repeated measurements (mean square of trials, MS_T). This means that the ICC_A provides more accurate predictions compared to the ICC_C. The ICC_A should be the choice of analysis for testing the reliability of instruments measuring continuous outcomes. In addition, the interpretation of ICC should be based on the confidence intervals, not on the ICC value.

There were significant differences between the predictions of ICC values from two and three repeated measurements. The predictions with three repeated measurements were more accurate compared to the prediction with two repeated measurements. This suggests that it is important to have at least three repeated measurements for each subject when testing the reliability of any instruments.

6.4 Recommendations

It is imperative that all medical instruments are accurate and precise. Otherwise, a failure may lead to critical medical errors. Therefore, there is a necessity for the proper evaluation of all medical instruments, and it is important to be sure that the appropriate statistical method has been used. As discussed in chapter 5, a method that is fool proof for a method comparison study is required, so the recommendation in this thesis will be based on the best available method found in this study.

6.4.1 Recommendation on analysis in a method comparison study

Most of the statistical analysis is complex and difficult to perform or interpret. This recommendation is intended to guide medical researchers in the analysis of a method comparison study. The proposed analysis is a simple step-by-step analysis to overcome some of the problems or limitations of the suggested statistical methods.

1. Test the reliability of the investigated instrument first. This assumes that the standard or referral instrument is already precise and accurate (otherwise the instrument will not be a standard). If the instrument is not reliable, then something needs to be done to ensure its precision. There is no reason to continue testing the accuracy of imprecise instruments. The recommended statistical method for analysis is the ICC_A (intra-class correlation coefficient for absolute agreement) with at least three repeated measurements. The CIs should be reported and the conclusion on reliability must be based on the confidence interval values.
2. Once the instrument is reliable, then the correlation between the measurements from the tested instrument and the standard instrument should be checked. If there is no strong linear relationship between the tested and standard

instruments, then it is impossible to detect any agreement between the two instruments (i.e. confirm disagreement).

3. If there is a strong correlation between the tested and standard instruments, then there is a possibility that the two instruments will be in agreement. To test for agreement:
    i. Set the acceptable clinical difference for the variable (e.g. 10mmHg for SBP and 0.5kg for body weight).
    ii. Run the Bland-Altman analysis and the agreement model. If no agreement is found in both the Bland-Altman analysis and the agreement model, there is truly no agreement between the instruments. When the Bland-Altman analysis resulted in the agreement of instruments, it is very likely that the instruments are truly in agreement. If no agreement is found in the Bland-Altman analysis, but agreement is found in the agreement model, the histogram of the differences should then be plotted, the proportion of differences in the dataset should be identified, and whether the proportion of differences is acceptable should be determined. This is based on clinical judgement, and the acceptable proportion of differences will depend on the type of instrument and the clinical setting where the instrument will be applied (i.e. critical care or health screening centre). These steps are summarised in Table 6.3 (Table of Agreement).

Table 6.3: Table of Agreement

		Bland-Altman LoA	
		Agreement	**No agreement**
Agreement Model	**No Agreement**	**YES (GOOD AGREEMENT)**	**NO (POOR AGREEMENT)**
	Agreement	**YES (GOOD AGREEMENT)**	**UNCERTAIN (Histogram & Clinical judgement)**

6.4.2 Recommendations on sample size

Another issue discussed in this thesis was on the sample size in the analysis of agreement. The sample size estimation for the regression analysis can be calculated from a published formula (J. Cohen, 1977). However the estimation of sample size for the Bland-Altman analysis and the ICC are based on precision.

Based on the findings in this study, it is recommended that a sample size of at least 100 is required for the Bland-Altman analysis for the prediction to be precise, and a larger sample size will result in a more precise prediction. A samples size of 100 is also a reasonably safe sample size for the analysis of reliability using $ICC_{A.}$ However, a sample size of more than 200 will be a waste of resources because the precision of the prediction will not change significantly if the sample size is more than 200.

6.5 Contribution of study

This study has contributed to two main areas. This includes a contribution to the medical research and a contribution to the community.

6.5.1 Contribution to medical research

6.5.1.1 Systematic reviews

This study presents two systematic reviews that provide evidence for the most popular statistical method used in method comparison studies.

1. The first systematic review is the first study specifically designed to retrieve information on statistical methods used to test for the agreement of instruments measuring the same continuous variables in medical literature. The Bland-Altman method was found to be the most cited paper in statistics (Ryan & Woodall, 2005). This leads one to think that this method is the most popular method used to test for agreement. However, citations do not imply that this method has been applied in research. This study provides supporting evidence that confirms the anecdotal claim that the Bland-Altman method is the most popular method used to assess agreement.
2. The second systematic review is the first ever systematic review conducted to identify the statistical methods used to assess the reliability of medical instruments with a continuous outcome. This study showed that the ICC is the most popular method that has been used to assess the reliability of medical instruments.

The findings from both reviews also revealed that most medical researchers only focused on a single factor in determining the quality of instruments. The agreement and reliability parameters were not assessed together in a single validation study. Most importantly, this review highlighted the use of an inappropriate statistical method in the analysis of agreement and reliability.

6.5.1.2 A few issues highlighted in a method comparison study in medicine

The extensive analysis and comparison of different statistical methods in this study has identified a few issues related to method comparison studies. These include:

1. The importance of testing both agreement and reliability in a method comparison study.
2. The need for an appropriate sample size for analysis by a method comparison study, especially in analysis using the Bland-Altman method and ICC.
3. The overestimation of bias in the Bland-Altman analysis, and the proposed regression analysis of the Bland-Altman Plot will not resolve the problem with an overestimation of bias due to proportional bias.
4. Findings from this study also raised the question of whether the confidence interval for the limits of agreement for the Bland-Altman analysis is really needed.
5. This study also emphasised that the ICC for absolute agreement (ICC_A) is the method of choice for the analysis of reliability for medical instruments measuring continuous variables.

6.5.1.3 Recommendation for analysis in a method comparison study

The most important contribution of this study is the recommendation of how to perform the analysis for a method comparison study, including the recommendation on statistical methods to be used for the analysis of agreement and reliability, and appropriate sample sizes. The flow of the proposed analysis is shown in Figure 6.1.

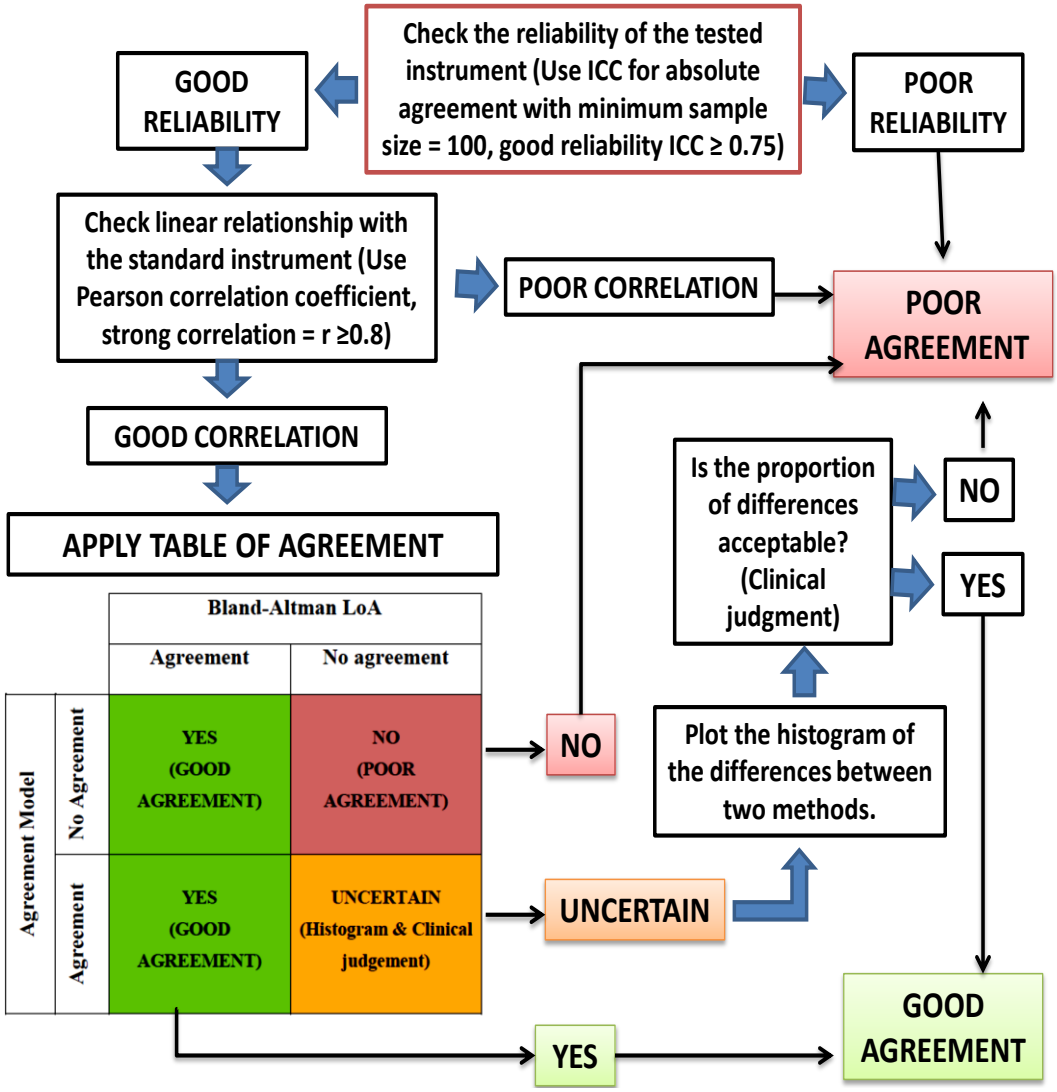


Figure 6.1: Proposed flow chart for the analysis in method comparison study

6.5.2 Community

The community health screening programme arranged by the researcher in the local village for the purpose of data collection has contributed to the local community. Most of the members in the community did not attend any regular health screens, and only seek medical help when they are really unwell. Therefore, the opportunity created was used by the villagers to have a general medical consultation with the researcher.

It does not matter how simple the variable is, but when it involves primary data collection in a community, it will always be challenging. Sometimes, things that are expected to be easy will turn out to be difficult, while things that are considered to be difficult to perform there will always be a solution or a way to make it easier. Nonetheless, in medical research, the needs of a participant as a patient should always be a priority.

6.6 Future work

There are a number of areas in which this research project can be extended in the future:

1. Recently, Bland and Altman republished their classic article (Bland & Altman, 1986) in the *International Journal of Nursing Studies* (Bland & Altman, 2010). The systematic review conducted earlier in Chapter 2 can be repeated for the studies published after the year of 2010 to identify whether this article (Bland & Altman, 2010) has an impact on the awareness or knowledge in a method comparison study, and also to identify if there is any reduction in the number of inappropriate methods used in the analysis of a method comparison study.
2. The findings from the two systematic reviews conducted in chapter 2 can be a surrogate measure of the knowledge among medical researchers in a method comparison study. The application of inappropriate statistical methods in the

analysis suggests that their knowledge is still lacking. However, a direct measurement of the knowledge and their practice in the method comparison study should be conducted to confirm this.

3. This study only reviewed the most commonly used methods in medicine, so further studies are required to test the suitability of other methods that have been applied or proposed to be applied in the analysis of a method comparison study, such as the Passing and Bablok method (Bilić-Zulle, 2011). Data collection from population of a multiple setting (normal population and hospital setting) will help to improve the range of data collected (covering the extreme abnormal values), and improve the representative of the extreme values in the dataset.
4. This study found that testing the proportional bias using the regression analysis of the Bland-Altman plot did not resolve the overestimation problem in Bland-Altman analysis. Further analysis with various ranges of variables and ranges of bias is required to identify other ways of handling this issue, and to determine the impact of this overestimation of bias in clinical practice.
5. The tendency of the Bland-Altman analysis to overestimate bias also suggests that further studies are required to confirm whether the reporting of confidence intervals for the Bland-Altman analysis is really necessary.

6.7 Summary

In this chapter, the recommended flow of analysis of agreement and reliability in a method comparison study is presented. Although there is an overestimation of bias in the Bland-Altman analysis, this does not suggest that this method should be abandoned. All statistical methods have their weaknesses**.** The issue of detecting inconsistent error has been avoided by ensuring the reliability of the tested instrument. The proposed flow analysis (in combination with the agreement model and histogram of differences) is planned to overcome some of the limitations of the Bland-Altman method**.**

This chapter also recalls the contribution of this study to the medical area and to the community. This includes findings from the systematic reviews and recommendations on the analysis in a method comparison study. The suggestion for future work is offered at the end of the thesis. These are repeated systematic reviews for the studies published after the year 2010, the direct assessment of knowledge on method comparison studies, studies of other statistical methods, further studies to solve problems with the overestimation of bias in the Bland-Altman analysis, and the need for confidence intervals in the Bland-Altman limits of agreement.

Finally, the inappropriate analysis in the method comparison study is a cause for concern in the medical field and cannot be ignored. It is important for medical researchers and clinicians from all specialties to be aware of this issue because inappropriate statistical analyses will lead to inappropriate conclusions, thus jeopardising the quality of the evidence, which may, in turn, influence the quality of care given to the patients.

References

Ageberg, E., Flenhagen, J., & Ljung, J. (2007). Test-retest reliability of knee kinesthesia in healthy adults. *BMC Musculoskeletal Disorders, 8*.

Ahn, Y., & Garruto, R. M. (2008). Estimations of body surface area in newborns. *Acta Paediatrica (Oslo, Norway: 1992), 97*(3), 366-370.

Ahn, Y., Kwon, E., Shim, J. E., Park, M. K., Joo, Y., Kimm, K., . . . Kim, D. H. (2007). Validation and reproducibility of food frequency questionnaire for Korean genome epidemiologic study. *European Journal Of Clinical Nutrition, 61*(12), 1435-1441.

Allen, Ryan, Wallace, Lance, Larson, Timothy, Sheppard, Lianne, & Liu, Lee-Jane Sally. (2007). Evaluation of the recursive model approach for estimating particulate matter infiltration efficiencies using continuous light scattering data. *Journal Of Exposure Science & Environmental Epidemiology, 17*(5), 468-477.

Altman, D.G., & Bland, J.M. (1983). Measurement in Medicine: the analysis of method comparison studies. *The Statistician, 32*, 307-317.

Altman, D.G., & Bland, J.M. (2005). Standard deviations and standard errors. *British Medical Journal, 331*(7521), 903. doi: 10.1136/bmj.331.7521.903

Anderst, William, Zauel, Roger, Bishop, Jennifer, Demps, Erinn, & Tashman, Scott. (2009). Validation of three-dimensional model-based tibio-femoral tracking during running. *Medical Engineering & Physics, 31*(1), 10-16.

Andrieux, Pierre, Kilinc, Tamara, Perrin, Christian, & Campos-Gimenez, Esther. (2008). Simultaneous determination of free carnitine and total choline by liquid chromatography/mass spectrometry in infant formula and health-care products: single-laboratory validation. *Journal Of AOAC International, 91*(4), 777-785.

Antona, B., Barra, F., Barrio, A., Gonzalez, E., & Sanchez, I. (2007). Validity and repeatability of a new test for aniseikonia. *Investigative Ophthalmology and Visual Science, 48*(1), 58-62.

Arjomand, Lari H (2002). Business Statistics: Sampling Distribution of the Mean (Lecture 7). *Sampling Distribution of the Mean (Lecture 7).* Retrieved 28 April, 2011, from http://business.clayton.edu/arjomand/business/l7.html

Bahareh, Amirkalali., Saeed, Hosseini. , Ramin, Heshmat., & Bagher, Larijani. (2008). Comparison of Harris Benedict and Mifflin-St Jeor equations with indirect calorimetry in evaluating resting energy expenditure. *Indian Journal of Medical Sciences, 62*(7), 283-290.

Barthelemy, Jonathan, Gregor, Ladislav, Krejci, Ivo, Wataha, John, & Bouillaguet, Serge. (2009). Accuracy of electronic apex locater-controlled handpieces. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology, 107*(3), 437-441.

Bashiria, Mahdi, & Moslemia, Amir. (2011). A robust moving average iterative weighting method to analyze the effect of outliers on the response surface design. *International Journal of Industrial Engineering Computations, 2*, 851-862.

Batterham, Alan M. (2004). Commentary on bias in Bland-Altman but not regression validity analyses. *Sportscience, 8*, 47-44.

Baxter, P. (2005). Invalid measurement validity. *Developmental Medicine & Child Neurology, 47*, 291.

Bilić-Zulle, Lidija. (2011). Comparison of methods: Passing and Bablok regression. *Biochemia Medica, 21*(1), 49-52.

Bland, J. M. (1995). *An Introduction to Medical Statistics* (2nd ed.). Oxford: Oxford University Press.

Bland, J. M., & Altman, D. G. (1990). A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Computers in Biology and Medicine, 20*(5), 337-340.

Bland, J.M. (2004). Sample size for a study of agreement between two methods of measurement. Retrieved 27 September, 2011, from http://www-users.york.ac.uk/~mb55/meas/sizemeth.htm

Bland, J.M., & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, i*, 307-310.

Bland, J.M., & Altman, D.G. (1987). Statistical methods for assessing agreement between two methods of clinical measurement. *Biochimica Clinica, 11*, 399-404.

Bland, J.M., & Altman, D.G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135-160.

Bland, J.M., & Altman, D.G. (2003). Applying the right statistics: analyses of measurement studies. *Ultrasound in Obstetrics & Gynecology, 22*(1), 85-93. doi: 10.1002/uog.122

Bland, J.M., & Altman, D.G. (2007). Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics, 17*(4), 571-582. doi: 10.1080/10543400701329422

Bland, J.M., & Altman, D.G. (2010). Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies, 47*, 931-936.

Bland, J.M., & Altman, D.G. . (2012). Agreed Statistics: Measurement Method Comparison. *Anesthesiology, 116*(1), 182-185.

Boyles, Sarah Hamilton, Edwards, S. Renee, Gregory, W. Thomas, Denman, Mary Anna, & Clark, Amanda L. (2007). Validating a clinical measure of levator hiatus size. *American Journal of Obstetrics and Gynecology, 196*(2), 174.e171-174.e174.

Bruton, A., Conway, J.H., & Holgate, S.T. (2000). Reliability: What is it, and how is it measured? *Physiotherapy, 86*(2), 94-99.

Burke, M. J. , & Whelan, M. V. (1987). The accuracy and reliability of commercial heart rate monitors. *British Journal of Sports Medicine, 21*(1), 29-32.

Camara, Oscar, Schnabel, Julia A., Ridgway, Gerard R., Crum, William R., Douiri, Abdel, Scahill, Rachael I., . . . Fox, Nick C. (2008). Accuracy assessment of global and local atrophy measurement techniques with realistic simulated longitudinal Alzheimer's disease images. *NeuroImage, 42*(2), 696-709.

Chambers, Matthew S. (2008). Chapter 6: Assessing the Assumptions of the Regression Model. Retrieved 29 April 2011, 2011, from http://pages.towson.edu/mchamber/chapter6eco306.pdf

Chan, Y H. (2003). Biostatistics 104: Correlational Analysis. *Singapore Medical Journal, 44*(12), 614-619.

Chovel Cuervo, ML, Sterling, AL , Abreu Nicot, I , García Rodríguez, M , & Rodríguez García, O. (2008). Validation of a new alternative for determining in vitro potency in vaccines containing Hepatitis B from two different manufacturers. *Biologicals, 36*(6), 375-382.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (Revised ed.). New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Second ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, M., Boyle, E., Delaney, C., & Shaw, J. (2006). A comparison of blood glucose meters in Australia *Diabetes Research and Clinical Practice, 71,* 113-118.

Cohen, M. D., & Jennings, S. G. (2002). Agreement and reproducibility of subjective methods of measuring faculty time distribution. *Academic Radiology, 9*(10), 1201-1208.

Cuker, A., Ptashkin, B., Konkle, B. A., Pipe, S. W., Whinna, H. C., Zheng, X. L., . . . Pollak, E. S. (2009). Interlaboratory agreement in the monitoring of unfractionated heparin using the anti-factor Xa-correlated activated partial thromboplastin time. *Journal of Thrombosis and Haemostasis, 7*(1), 80-86.

Daly, Leslie. E., & Bourke, Geoffrey. J. (2000). *Interpretation and Use of Medical Statistics* (5th ed.). Oxford: Blackwell Science Ltd.

de Vet, Henrica C.W., Terwee, Caroline B. , & Bouter, Lex M. (2003a). Clinimetrics versus psychometrics: two sides of the same coin. *Journal of Clinical Epidemiology, 56,* 1146-1147.

de Vet, Henrica C.W., Terwee, Caroline B. , & Bouter, Lex M. (2003b). Current challenges in clinimetrics. *Journal of Clinical Epidemiology, 56*(12), 1137-1141. doi: S0895435603003573

de Vet, Henrica C.W., Terwee, Caroline B., Knol, Dirk L., & Bouter, Lex M. . (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology, 59*, 1033-1039.

Di Noia, Jennifer, & Contento, Isobel R. (2009). Use of a brief food frequency questionnaire for estimating daily number of servings of fruits and vegetables in a minority adolescent population. *Journal of the American Dietetic Association, 109*(10), 1785-1789.

Digital Scales Company: Hanson H926 mechanical bathroom scale. (2011). Retrieved 23 February 2011, from http://www.digital-scales-company.com/index.php?main_page=product_info&cPath=87&products_id=478&zenid=e30f2e0e33dc1d535b053448f524e2aa

Doros, Gheorghe, & Lew, Robert. (2010). Design Based on Intra-Class Correlation Coefficients. *American Journal of Biostatistics, 6*(1), 1-8.

Dunsky, Eliot H, Goldstein, Marc F, Dvorin, Donald J, Belecanech, George A, Haralabatos, Irene C, Gordon, Nancy D, & Moday, Heather J. (2005). Understanding Asthma (4th ed.). Philadelphia: The Asthma Center Education and Research Fund Manual.

Essack, Y., Hoffman, M., Rensburg, M., Van Wyk, J., Meyer, C.S., & Erasmus, R. (2009). A comparison of five glucometers in South Africa. *Journal of Endocrinology, Metabolism and Diabetes of South Africa, 14*(2), 102-105.

Fay, M. P. (2005). Random marginal agreement coefficients: rethinking the adjustment for chance when measuring agreement. *Biostatistics, 6*(1), 171-180.

Feinstein, A.R. . (1987). *Clinimetrics*. New Haven: Yale University Press.

Fisher, Joan. (1978). *R. A. Fisher: The Life of a Scientist*. New York: Wiley.

Fisher, Ronald. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Hacihaliloglu, Ilker, Abugharbieh, Rafeef, Hodgson, Antony J., & Rohling, Robert N. (2009). Bone surface localization in ultrasound using image phase-based features. *Ultrasound in Medicine & Biology, 35*(9), 1475-1487.

Hamilton, C , & Stamey, J. (2007). Using Bland Altman to assess agreement between two medical devices – don't forget the confidence intervals! *Journal of Clinical Monitoring and Computing 21*, 331-333.

Hanbazaza, Sarfinaz M. , & Mansoor, Ibrahim. . (2012). Accuracy evaluation of point-of-care glucose analyzers in the Saudi market. *Saudi Medical Journal, 33*(1), 91-92.

Haynes, S.N., Richard, D.C.S., & Kubany, E.S. (1995). Content validity in psychological assessment: a functional approach to concepts and methods. *Psychological Assessment, 7*, 238-247.

Hibbert, D. Brynn. (2007). Systematic errors in analytical measurement results. *Journal of Chromatography A,, 1158*, 25-32.

Hof, I, Arbab-Zadeh, A, Dong, J, Scherr, D, Chilukuri, K, & Calkins, H (2008). Validation of a simplified method to determine left atrial volume by computed tomography in patients with atrial fibrillation. *The American Journal of Cardiology, 102*(11), 1567-1570.

Holzinger, U., Warszawska, J., Kitzberger, R., Herkner, H., Metnitz, P. G., & Madl, C. (2009). Impact of shock requiring norepinephrine on the accuracy and reliability of subcutaneous continuous glucose monitoring. *Intensive Care Med, 35*(8), 1383-1389. doi: 10.1007/s00134-009-1471-y

Hopkins, W.G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine, 30*(1), 1-15.

Hopkins, W.G. (2004). Bias in Bland-Altman but not regression validity analyses. *Sportscience, 8*, 42-46.

Hunter, D., & Pierscionek, B.K. . (2007). Children, Gillick competency and consent for involvement in research. *Journal of Medical Ethics, 33*(11), 659-662.

Innes, E., & Straker, L. . (1999). Validity of work-related assessments. *Work, 13*, 125-152.

IR Thermometer (DT-8806H). (2009). *Product description.* Retrieved 1 March 2009, from http://ysun-amanda.en.made-in-china.com/product/QMcJhuaCJHWz/China-IR-Thermometer-DT-8806H-.html

Jaffrin, Michel Y., & Morel, Hélène. (2008). Body fluid volumes measurements by impedance: A review of bioimpedance spectroscopy (BIS) and bioimpedance analysis (BIA) methods. *Medical Engineering & Physics, 30*(10), 1257-1269.

JCGM. (2008). *Evaluation of measurement data - Guide to the Expression of Uncertainty in Measurement (ISO)*. Geneva: Joint Committee for Guides in Metrology Retrieved from http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf.

Kirkwood, Betty R. (2000). *Medical Statistics*. Oxford: Blackwell Science Ltd.

Klang District Office. (2011). Retrieved 15 February 2011, from http://www.Selangor.gov.my/klang

Kottner, Jan, Audigé, Laurent, Brorson, Stig, Donner, Allan, Gajewski, Byron J., Hróbjartsson, Asbjørn, . . . Streiner, David L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology, 64*(1), 96-106.

Lambert, H.C., Gisel, E.G., & Wood-Dauphinee, S. (2002). The Functional Assessment of Dysphagia: Psychometric Standards. *Physical & Occupational Therapy in Geriatrics, 19*, 1-14.

Larsen, Pia Veldt. (2008). Master of Applied Statistics: Regression and analysis of variance. Retrieved 29 April 2011, from http://statmaster.sdu.dk/courses/st111/module04/index.html

Lee, J., Koh, D., & Ong, C. N. (1989). Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Computers in Biology and Medicine, 19*(1), 61-70.

Lin, L. I. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine, 19*(2), 255-270. doi: 10.1002/(SICI)1097-0258(20000130)19:2<255::AID-SIM293>3.0.CO;2-8

Lowry, Richard. (2012). Concepts & Applications of Inferential Statistics. *Chapter 17. One-Way Analysis of Covariance for Independent Samples. Part 3.* Retrieved 20 October, 2012, from http://vassarstats.net/textbook/ch17pt3.html

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: a critical review. *Clinical and Experimental Pharmacology and Physiology, 29*(7), 527-536. doi: 3686 [pii]

Ludbrook, John. (2010). Confidence in Altman-Bland plots: a critical review of the method of differences. *Clinical and Experimental Pharmacology Physiology 37*, 143-149.

Ludbrook, John. . (2002). Statistical techniques for comparing measurers and methods of measurement: A critical review. *Clinical and Experimental Pharmacology and Physiology, 29*(7), 527-536. doi: 10.1046/j.1440-1681.2002.03686.x

Luiz, Ronir Raggio , Costa, Antonio José Leal , Kale, Pauline Lorena , & Werneck, Guilherme L (2003). Assessment of agreement of a quantitative variable: a new graphical approach. *Journal of Clinical Epidemiology, 56*(10), 963-967.

Luiz, Ronir Raggio , & Szklo, M. (2005). More than one statistical strategy to assess agreement of quantitative measurements may usefully be reported. *Journal of Clinical Epidemiology, 58*(4), 215-216.

Maksymowych, W. P., Dhillon, S. S., Park, R., Salonen, D., Inman, R. D., & Lambert, R. G. W. (2007). Validation of the spondylarthritis research consortium of Canada magnetic resonance imaging spinal inflammation index: Is it necessary to score the entire spine? *Arthritis Care and Research, 57*(3), 501-507.

Mårtensson, Mattias., Winter, Reidar., Cederlund, Kerstin. , Ripsweden, Jonaz. , Mir-Akbari, Habib., Nowak, Jacek., & Brodin, Lars-Åke. (2008). Assessment of left ventricular volumes using simplified 3-D echocardiography and computed tomography – a phantom and clinical study. *Cardiovascular Ultrasound 2008, 6:26, 6*(26).

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

McPhillips-Tangum, C. A., Aubert, R., Bailey, C. M., & Koplan, J. P. (1997). Measuring pediatric immunization status in a managed care organization: agreement between medical charts and parent telephone interviews. *HMO Practice, 11*(3), 104-110.

Milias, G. A. , Antonopoulou, S. , & Anthanasopoulos, S. . (2008). Development, reliability and validity of a new motorized isometric dynamometer for measuring strength characteristics of elbow flexor muscles. *Journal of Medical Engineering & Technology, 32*(1), 66-72.

Mini-Wright Standard. (2011). Retrieved 23 February 2011, from http://www.clement-clarke.com/products/mini-wright-standard

Moher, D., Liberati, A., Tetzlaff, J., Altman D.G., &The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. *PLoS Med,* 6(6): e1000097. doi:10.1371/journal.pmed1000097.

Motulsky, H.J. (2007). *Prism 5 Statistics Guide*. San Diego CA: Graph Pad Software Inc. Retrieved from www.graphpad.com.

Muller, R., & Buttner, P. (1994). A critical discussion of intraclass correlation coefficients. *Stat Med, 13*(23-24), 2465-2476.

Mündermann, A, Dyrby, CO, & Andriacchi, TP (2008). A comparison of measuring mechanical axis alignment using three-dimensional position capture with skin markers and radiographic measurements in patients with bilateral medial compartment knee osteoarthritis. *The Knee, 15*(6), 480-485.

Naidu, S.H., Panchik, D, & Chinchilli, V.M. (2009). Development and validation of the hand assessment tool. *Journal of Hand Therapy, 22*(3), 250-257.

National Weights and Measures Laboratory. (2003). *Consultation on effect of non-automatic weighing instruments Regulations 2000 SI 2000/3236 on medical weighing instruments put Into use after 1 january 2003.* Retrieved 21 January, 2011, from http://www.nmodoitonline.bis.gov.uk/Docs/Legislation/NAWI/Consult%20effects%20NAWI%20on%20medical%20instruments.pdf

Nazir, Z., Razaq, S., Mir, S., Anwar, M., Al Mawlawi, G., Sajad, M., . . . Taylor, R.S. (2005). Revisiting the accuracy of peak flow meters: a double-blind study using formal methods of agreement. *Respiratory Medicine, 99*, 592-595.

Neveu, D., Aubas, P., Seguret, F., Kramar, A., & Dujols, P. (2006). Measuring agreement for ordered ratings in 3 x 3 tables. *Methods of Information in Medicine, 45*(5), 541-547. doi: 06050541

NHF. (2009). *Guide to management of hypertension 2008: Assessing and managing raised blood pressure in adults*. Australia: National Heart Foundation of Australia Retrieved from http://www.heartfoundation.org.au.

NICE. (2011). National Institute for Health and Clinical Excellence UK clinical guideline 127 *Hypertension: Clinical management of primary hypertension in adults.* Retrieved 26 April, 2012, from www.nice.org.uk/guidance/CG127

Omron Instruction Manual. (2009). *Omron HEM-907KL.* Retrieved 23 February 2009, from http://www.omronhealthcare.com/media/uploads/hem-907xl_im.pdf

OpenEpi, Version 2, open source calculator--PowerMean.). Retrieved 5 May, 2012, from http://www.openepi.com/OE2.3/Power/PowerMean.htm

Oztuna, Derya, Elhan, Atilla Halil, & Tuccar, Ersöz. (2006). Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions. *Turkish Journal of Medical Sciences, 36*(3), 171-175.

Passing, H., & Bablok, W. . (1983). A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part I. *Journal of Clinical Chemistry & Clinical Biochemistry, 21*, 709-720.

Pearson, Karl (1905). "Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson". A rejoinder. *Biometrika*, 4(1), 169-212.

Pekeliling Perkhidmatan Bilangan 3 (2003). *Pemeriksaan Kesihatan Bagi Pegawai Perkhidmatan Awam.* Retrieved 8 August, 2012, from http://www.jpa.gov.my/pekeliling/pp03/bil03/kesihatan.pdf.

Pennsylvania State University online course (2013). *STAT 897D Applied Data Mining and Statistical Learning. Lesson 1: Introduction to Data Mining*. Retrieved 3 July, 2013, from https://onlinecourses.science.psu.edu/stat857/node/30.

Pickering, Thomas G, Hall, John E , Appel, Lawrence J. , Falkner, Bonita E. , Graves, John , Hill, Martha N. , . . . Roccella, Edward J. . (2005). Recommendations for Blood Pressure Measurement in Humans and Experimental Animals. *Hypertension, 45*, 142-161.

piCO+ Smokerlyzer User Manual. (2006). Bedfont Scientific.

Pini, Claudio , Natalizi, Anna , Gerosa, Pietro Francesco , Frigerio, Mauro , Omboni, Stefano , & Parati, Gianfranco (2008). Validation of the Artsana CS 410 automated blood pressure monitor in adults according to the International Protocol of the European Society of Hypertension. *Blood Pressure Monitoring 13*, 177–182.

Pini, Claudio , Pastori, Marco , Baccheschi, Jordan , Omboni, Stefano , & Parati, Gianfranco (2007). Validation of the Artsana CSI 610 automated blood pressure monitor in adults according to the International Protocol of the European Society of Hypertension. *Blood Pressure Monitoring, 12*, 179- 184.

Portal Rasmi Kampung Tradisional. (2012). *Kampung Teluk Gadong.* Retrieved 8 August, 2012, from http://ejkt.kpkt.gov.my/kgtrad/index.php?page=4&kod_kg=8618

Portney, L.G., & Watkins, M.P. (2000). *Foundations of clinical research: Applications to practice*. New Jersey: Prentice-Hall.

Quanjer PH, Lebowitz MD, Gregg I, Miller MR, & Pedersen OF. (1997). Peak expiratory flow: conclusions and recommendations of a Working Party of the European Respiratory Society. *European Respiratory Journal, 10*(24), 2s-8s.

Razali, N., Wah, Y.B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling test. *Journal of Statistical Modeling and Analytics*, 2(1), 21–33.

Reilly K, Munro J, Pandit S, Kress A, Walker C, & Pitto RP. (2007). Inter-observer validation study of quantitative CT-osteodensitometry in total knee arthroplasty. *Arch Orthop Trauma Surg, 127*(8), 729-731.

Reis, MF, Aniceto, P , Aguiar, P, Simao, F , & Segurado, S (2007). Quantification of urinary chorionic gonadotropin in spontaneous abortion of pre-clinically recognized pregnancy: method development and analytical validation. *International Journal of Hygiene and Environmental Health, 3-4*(210), 419-427.

Robinson, Joan L. , Jou, Hsing , & Spady, Donald W. (2005). Accuracy of parents in measuring body temperature with a tympanic thermometer. *BMC Family Practice, 6*(3). doi: 10.1186/1471-2296-6-3

Roche Accu-Chek Owner's Booklet. (2004). Retrieved 1 March 2009, from http://www.northcoastmed.com/pdf/manuals/advantage_userguide.pdf

Rosner, B. (2006). *Fundementals of Biostatistics* (6th ed.). Duxbury: Thomson Brooks/Cole.

Ryan, T.P., & Woodall, W.H. (2005). The Most-Cited Statistical Papers. *Journal of Applied Statistics, 32*(5), 461-474.

Satia, J.A., & Galanko, J.A. . (2007). Comparison of three methods of measuring dietary fat consumption by African-American adults. *Journal of the American Dietetic Association, 107*(5), 782-791.

Satia, J.A., Watters, J.L., & Galanko, J.A. (2009). Validation of an antioxidant nutrient questionnaire in whites and African Americans. *Journal of the American Dietetic Association, 109*(3), 502-508.e506.

Scales Galore: High capacity bathroom scales (2011). Retrieved 23 February 2011, from http://www.scalesgalore.com/pbath.htm#high

Shannon, Harriet, Gregson, Rachael, Stocks, Janet, Cole, Tim J., & Main, Eleanor. (2009). Repeatability of physiotherapy chest wall vibrations applied to spontaneously breathing adults. *Physiotherapy, 95*(1), 36-42.

Shoukri, M.M., & Pause, C.A. (1999). *Statistical Methods for Health Sciences* (Second ed.). Boca Raton, Florida.: CRC Press.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass Correlations: uses in assessing rater reliability *Psychological Bulletin, 86*(2), 420-428.

Shuaibi, AM, Sevenhuysen, G.P., & House, J.D. (2008). Validation of a food choice map with a 3-day food record and serum values to assess folate and vitamin B-12 intake in college-aged women. *Journal of the American Dietetic Association, 108*(12), 2041-2050.

Smith, Mark W., Ma, Jun., & Stafford, Randall S. (2010). Bar charts enhance Bland-Altman plots when value ranges are limited. *Journal of Clinical Epidemiology, 63*(2), 180-184.

Smith, T Paul. (2012). Biol 310 Research Design and Analysis. *Chapter 16. Multiple Regression Analysis (ANCOVA).* Retrieved 20 October, 2012, from http://www.csub.edu/~psmith3/Teaching/310-12.pdf

Streiner, D.L. (1996). Maintaining standards: differences between the standard deviation and standard error, and when to use each. *Canadian Journal of Psychiatry, 41*, 498-502.

Streiner, D.L., & Norman, G.R. (2003). *Health measurement scales. Apractical guide to their development and use* (Third ed.). Oxford: Oxford University Press.

Syed, Faiz I., Oza, Ashish L., Vanderby, Ray., Heiderscheit, Bryan., & Anderson, Paul. A. . (2007). A method to measure cervical spine motion over extended periods of time. *Spine, 32*(19), 2092-2098.

Take Control of Your Asthma. (2012). *Asthma Management Tools.* Retrieved 8 March, 2012, from http://www.lung.org/lung-disease/asthma/living-with-asthma/take-control-of-your-asthma/

Ten Boekel, Edwin, Böck, Martine, Vrielink, Gert-Jan, Liem, Robert, Hendriks, Henriet, & Kieviet, Wim de. (2007). Detection of shortened activated partial thromboplastin times: An evaluation of different commercial reagents. *Thrombosis Research, 121*(3), 361-367.

Traditional Village in Selangor. (2010). Retrieved 15 February 2011, from http://www.kampungtradisiselangor.com

Ugrinowitsch, Carlos, Fellingham, Gilbert W., & Ricard, Mark D. (2004). Limitations of Ordinary Least Squares Models in Analyzing Repeated Measures Data. *Medicine and Science in Sports and Exercise., 36*(12), 2144-2148.

University of Malaya Official Portal.). Retrieved 8 August, 2012, from http://www.um.edu.my/mainpage.php?module=Maklumat&kategori=51&id=255&papar=1

Weir, Joseph. P. (2005). Quantifying test-retest reliability using the Intraclass Correlation Coefficient and the SEM. *Journal of Strength and Conditioning Research,, 19*(1), 231-240.

Wikipedia. (2009). The Free Encyclopedia. *Correlation.* Retrieved 30 January 2009, from http://en.wikipedia.org/wiki/Correlation

Yarows, A. Steven. , Patel, Ketul. , & Brook, Robert. . (2001). Rapid oscillometric blood pressure measurement compared to conventional oscillometric measurement. *Blood Pressure Monitoring, 6*(3).

Zaki, R., Bulgiba, A., Ismail, R., & Ismail, N.A. (2012). Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review. *PLoS ONE, 7: e37908. doi:10.1371/journal.pone.0037908*(5). doi: e37908. doi:10.1371/journal.pone.0037908

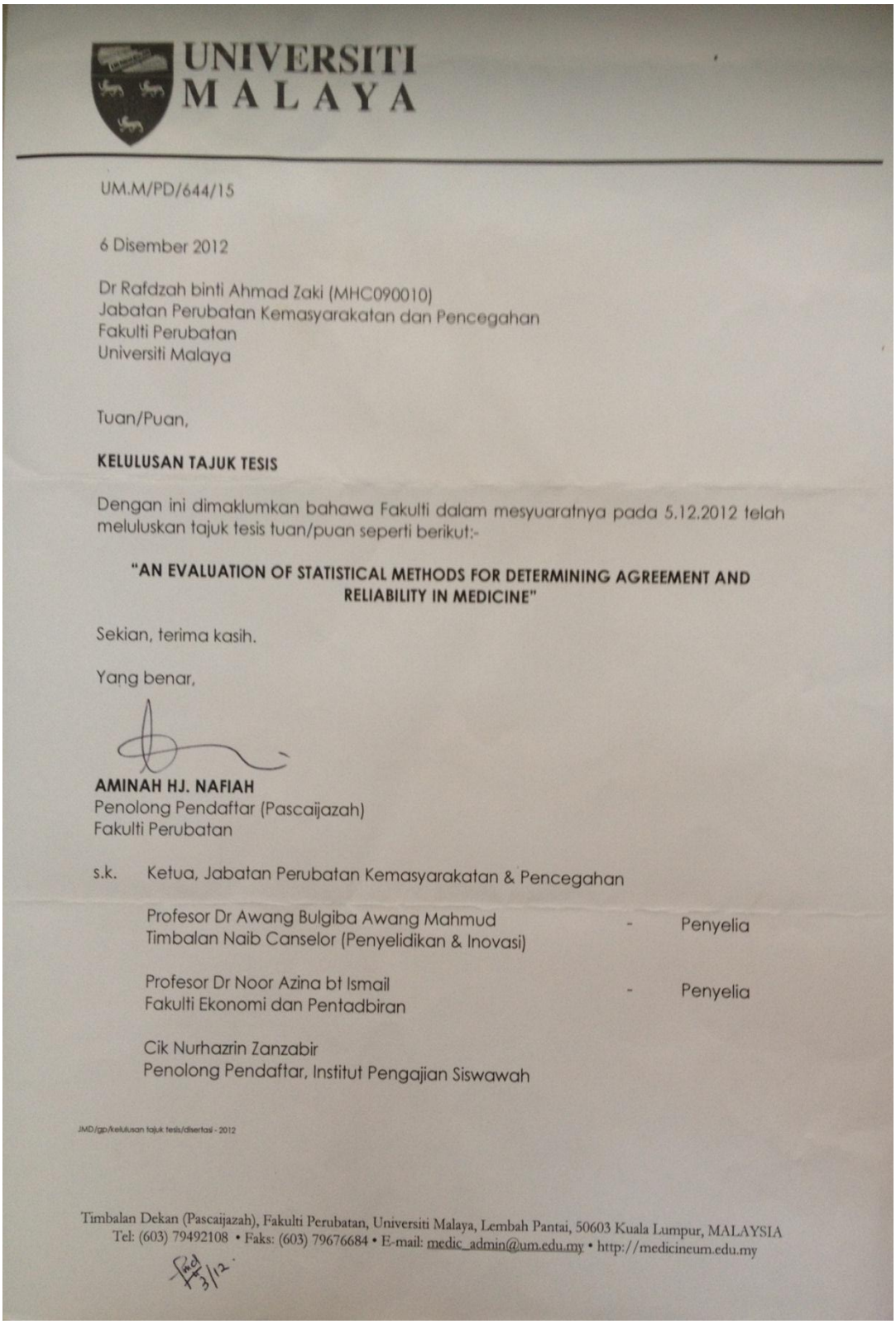
Zaki, Rafdzah, Bulgiba, Awang , Ismail, Roshidi, & Ismail, NoorAzina (2010, 29 August - 2 September). *A Review of Statistical Methods Used to Assess Agreement in Medicine.* Paper presented at the 31st Annual Conference of The International Society for Clinical Biostatistics (ISCB), Montpellier, France.

Zaki, Rafdzah, Bulgiba, Awang, Nordin, Noorhaire, Ismail, Roshidi, & Ismail, NoorAzina. (2010, 24-27November). *Knowledge on Methods of Validation Study In Medicine.* Paper presented at the 42th Asia Pasific Academic Consortium for Public Health Conference, Bali, Indonesia.

Zaki, Rafdzah, Nordin, Noorhaire, Bulgiba, Awang , & Ismail, NoorAzina (2010). *Assessing methods to determine reliability of medical instruments*. Paper presented at the 2nd International Conference on Quantitative Sciences and Its Applications (ICOQSIA), Penang, Malaysia.

Zar, Jerrold H. (2010). *Biostatistical Analysis* (5th ed.). New Jersey: Pearson Education International.

APPENDIX A: Topic approval



UNIVERSITI MALAYA

UM.M/PD/644/15

6 Disember 2012

Dr Rafdzah binti Ahmad Zaki (MHC090010)
Jabatan Perubatan Kemasyarakatan dan Pencegahan
Fakulti Perubatan
Universiti Malaya

Tuan/Puan,

KELULUSAN TAJUK TESIS

Dengan ini dimaklumkan bahawa Fakulti dalam mesyuaratnya pada 5.12.2012 telah meluluskan tajuk tesis tuan/puan seperti berikut:-

"AN EVALUATION OF STATISTICAL METHODS FOR DETERMINING AGREEMENT AND RELIABILITY IN MEDICINE"

Sekian, terima kasih.

Yang benar,

AMINAH HJ. NAFIAH
Penolong Pendaftar (Pascaijazah)
Fakulti Perubatan

s.k. Ketua, Jabatan Perubatan Kemasyarakatan & Pencegahan

Profesor Dr Awang Bulgiba Awang Mahmud - Penyelia
Timbalan Naib Canselor (Penyelidikan & Inovasi)

Profesor Dr Noor Azina bt Ismail - Penyelia
Fakulti Ekonomi dan Pentadbiran

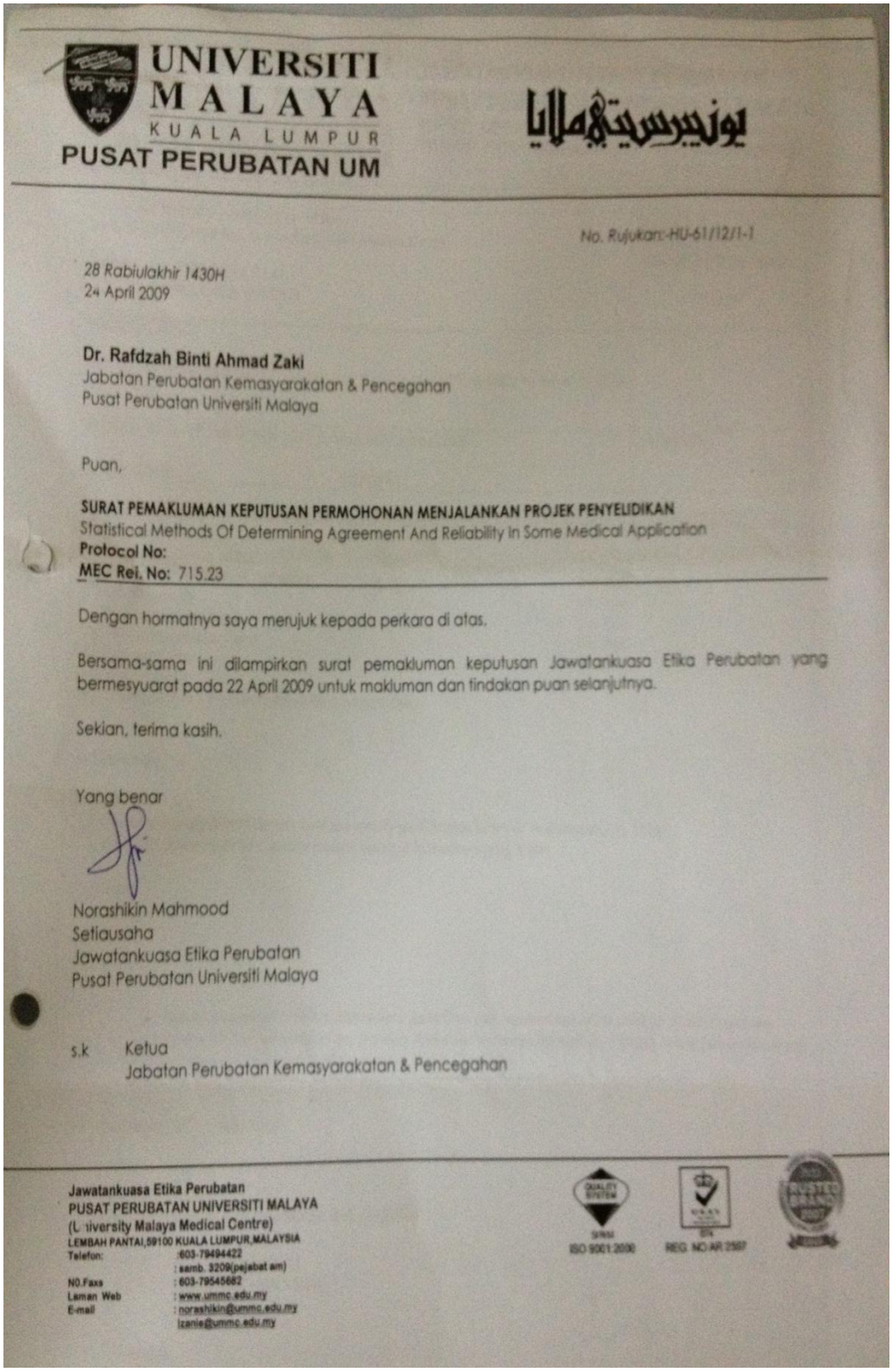
Cik Nurhazrin Zanzabir
Penolong Pendaftar, Institut Pengajian Siswawah

JMD/gp/Kelulusan tajuk tesis/disertasi - 2012

Timbalan Dekan (Pascaijazah), Fakulti Perubatan, Universiti Malaya, Lembah Pantai, 50603 Kuala Lumpur, MALAYSIA
Tel: (603) 79492108 • Faks: (603) 79676684 • E-mail: medic_admin@um.edu.my • http://medicineum.edu.my

3/12

APPENDIX B: Ethical approval



UNIVERSITI MALAYA
KUALA LUMPUR
PUSAT PERUBATAN UM

يونيورسيتي مالايا

No. Rujukan:-HU-61/12/1-1

28 Rabiulakhir 1430H
24 April 2009

Dr. Rafdzah Binti Ahmad Zaki
Jabatan Perubatan Kemasyarakatan & Pencegahan
Pusat Perubatan Universiti Malaya

Puan,

SURAT PEMAKLUMAN KEPUTUSAN PERMOHONAN MENJALANKAN PROJEK PENYELIDIKAN
Statistical Methods Of Determining Agreement And Reliability In Some Medical Application
Protocol No:
MEC Rei. No: 715.23

Dengan hormatnya saya merujuk kepada perkara di atas.

Bersama-sama ini dilampirkan surat pemakluman keputusan Jawatankuasa Etika Perubatan yang bermesyuarat pada 22 April 2009 untuk makluman dan tindakan puan selanjutnya.

Sekian, terima kasih.

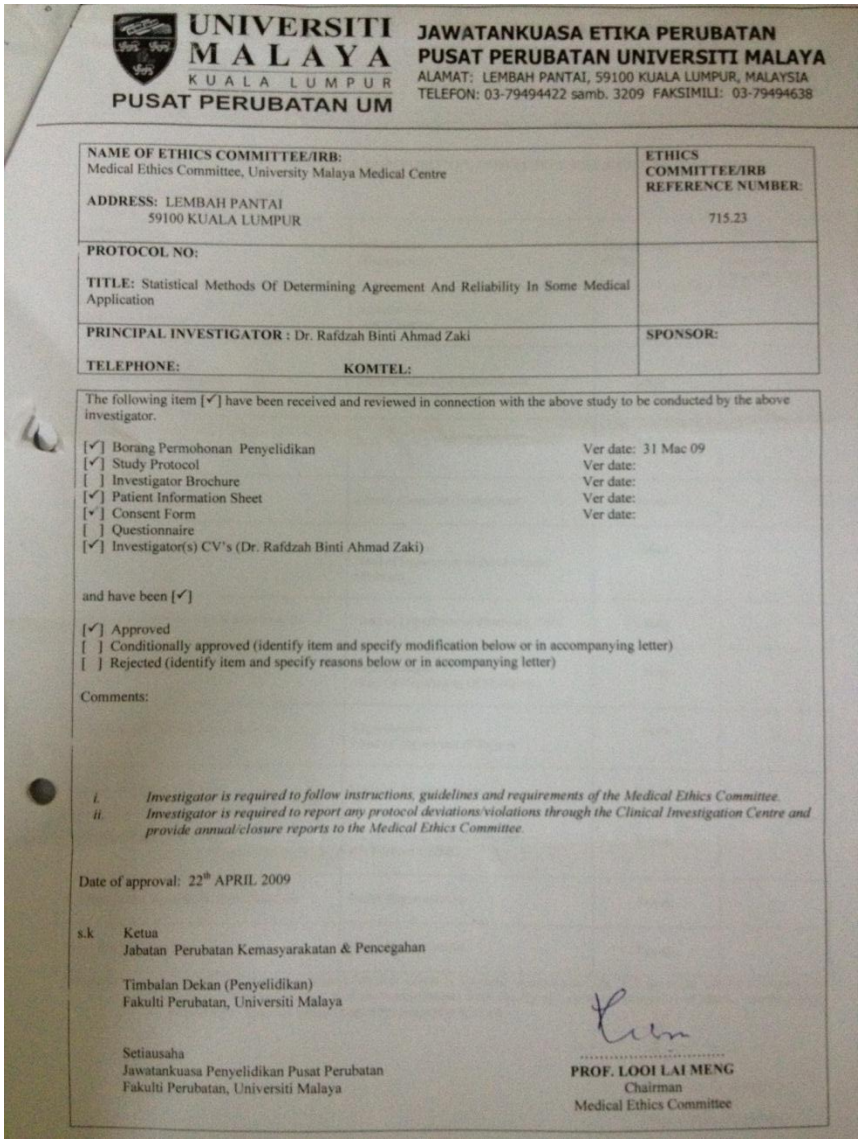
Yang benar

Norashikin Mahmood
Setiausaha
Jawatankuasa Etika Perubatan
Pusat Perubatan Universiti Malaya

s.k Ketua
Jabatan Perubatan Kemasyarakatan & Pencegahan

Jawatankuasa Etika Perubatan
PUSAT PERUBATAN UNIVERSITI MALAYA
(University Malaya Medical Centre)
LEMBAH PANTAI,59100 KUALA LUMPUR,MALAYSIA
Telefon: :603-79494422
: samb. 3209(pejabat am)
NO.Faks : 603-79545682
Laman Web : www.ummc.edu.my
E-mail : norashikin@ummc.edu.my
izanie@ummc.edu.my

ISO 9001:2000 REG. NO:AR 2567



UNIVERSITI MALAYA
KUALA LUMPUR
PUSAT PERUBATAN UM

JAWATANKUASA ETIKA PERUBATAN
PUSAT PERUBATAN UNIVERSITI MALAYA
ALAMAT: LEMBAH PANTAI, 59100 KUALA LUMPUR, MALAYSIA
TELEFON: 03-79494422 samb. 3209 FAKSIMILI: 03-79494638

NAME OF ETHICS COMMITTEE/IRB: Medical Ethics Committee, University Malaya Medical Centre **ADDRESS:** LEMBAH PANTAI 59100 KUALA LUMPUR	**ETHICS COMMITTEE/IRB REFERENCE NUMBER:** 715.23
PROTOCOL NO: **TITLE:** Statistical Methods Of Determining Agreement And Reliability In Some Medical Application	
PRINCIPAL INVESTIGATOR : Dr. Rafdzah Binti Ahmad Zaki **TELEPHONE:** **KOMTEL:**	**SPONSOR:**

The following item [✓] have been received and reviewed in connection with the above study to be conducted by the above investigator.

[✓] Borang Permohonan Penyelidikan	Ver date: 31 Mac 09
[✓] Study Protocol	Ver date:
[ ] Investigator Brochure	Ver date:
[✓] Patient Information Sheet	Ver date:
[✓] Consent Form	Ver date:
[ ] Questionnaire	
[✓] Investigator(s) CV's (Dr. Rafdzah Binti Ahmad Zaki)	

and have been [✓]

[✓] Approved
[ ] Conditionally approved (identify item and specify modification below or in accompanying letter)
[ ] Rejected (identify item and specify reasons below or in accompanying letter)

Comments:

i. *Investigator is required to follow instructions, guidelines and requirements of the Medical Ethics Committee.*
ii. *Investigator is required to report any protocol deviations/violations through the Clinical Investigation Centre and provide annual/closure reports to the Medical Ethics Committee.*

Date of approval: 22th APRIL 2009

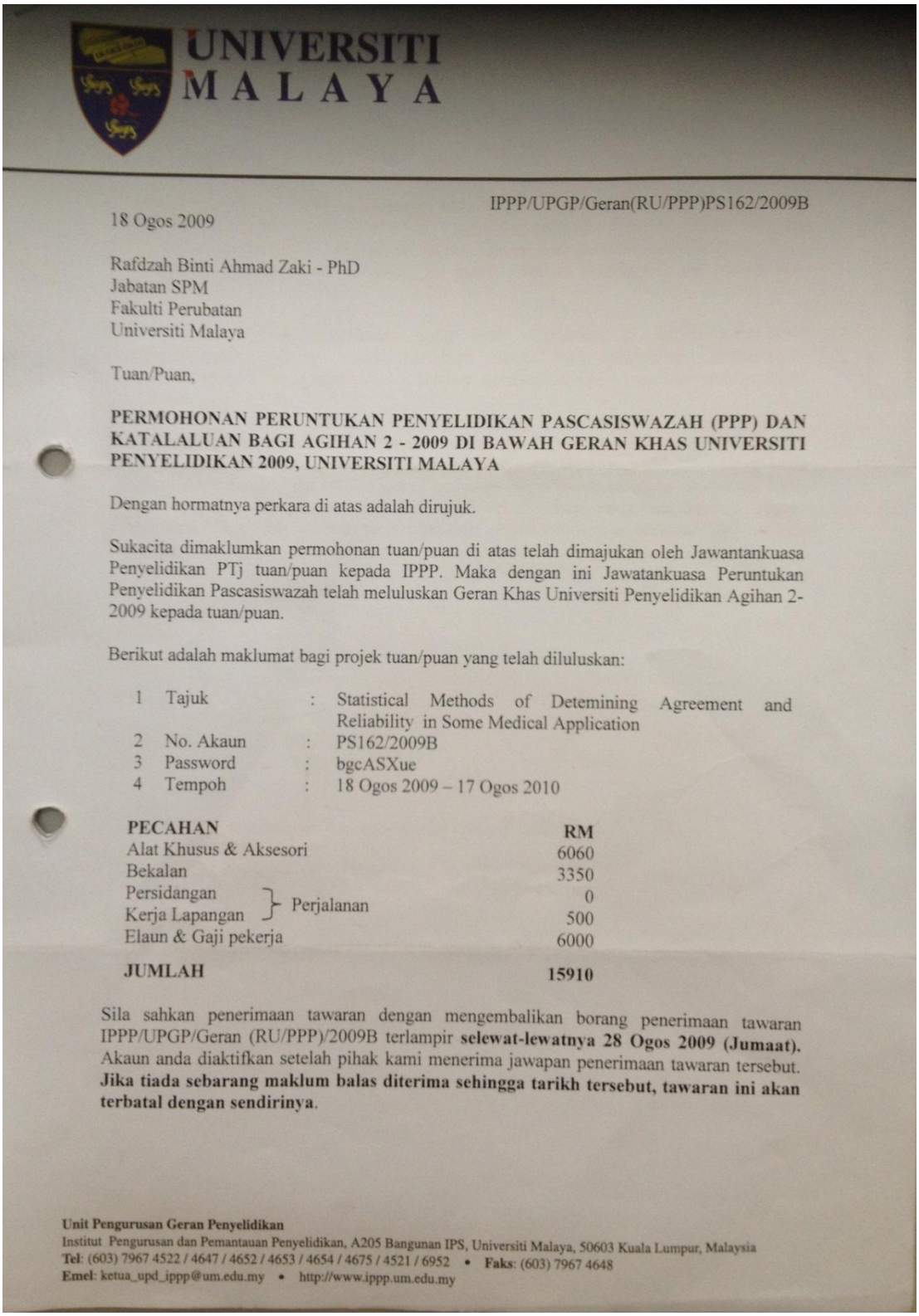
s.k Ketua
Jabatan Perubatan Kemasyarakatan & Pencegahan

Timbalan Dekan (Penyelidikan)
Fakulti Perubatan, Universiti Malaya

Setiausaha
Jawatankuasa Penyelidikan Pusat Perubatan
Fakulti Perubatan, Universiti Malaya

.....
PROF. LOOI LAI MENG
Chairman
Medical Ethics Committee

APPENDIX C: Funding approval



UNIVERSITI MALAYA

IPPP/UPGP/Geran(RU/PPP)PS162/2009B

18 Ogos 2009

Rafdzah Binti Ahmad Zaki - PhD
Jabatan SPM
Fakulti Perubatan
Universiti Malaya

Tuan/Puan,

PERMOHONAN PERUNTUKAN PENYELIDIKAN PASCASISWAZAH (PPP) DAN KATALALUAN BAGI AGIHAN 2 - 2009 DI BAWAH GERAN KHAS UNIVERSITI PENYELIDIKAN 2009, UNIVERSITI MALAYA

Dengan hormatnya perkara di atas adalah dirujuk.

Sukacita dimaklumkan permohonan tuan/puan di atas telah dimajukan oleh Jawantankuasa Penyelidikan PTj tuan/puan kepada IPPP. Maka dengan ini Jawatankuasa Peruntukan Penyelidikan Pascasiswazah telah meluluskan Geran Khas Universiti Penyelidikan Agihan 2-2009 kepada tuan/puan.

Berikut adalah maklumat bagi projek tuan/puan yang telah diluluskan:

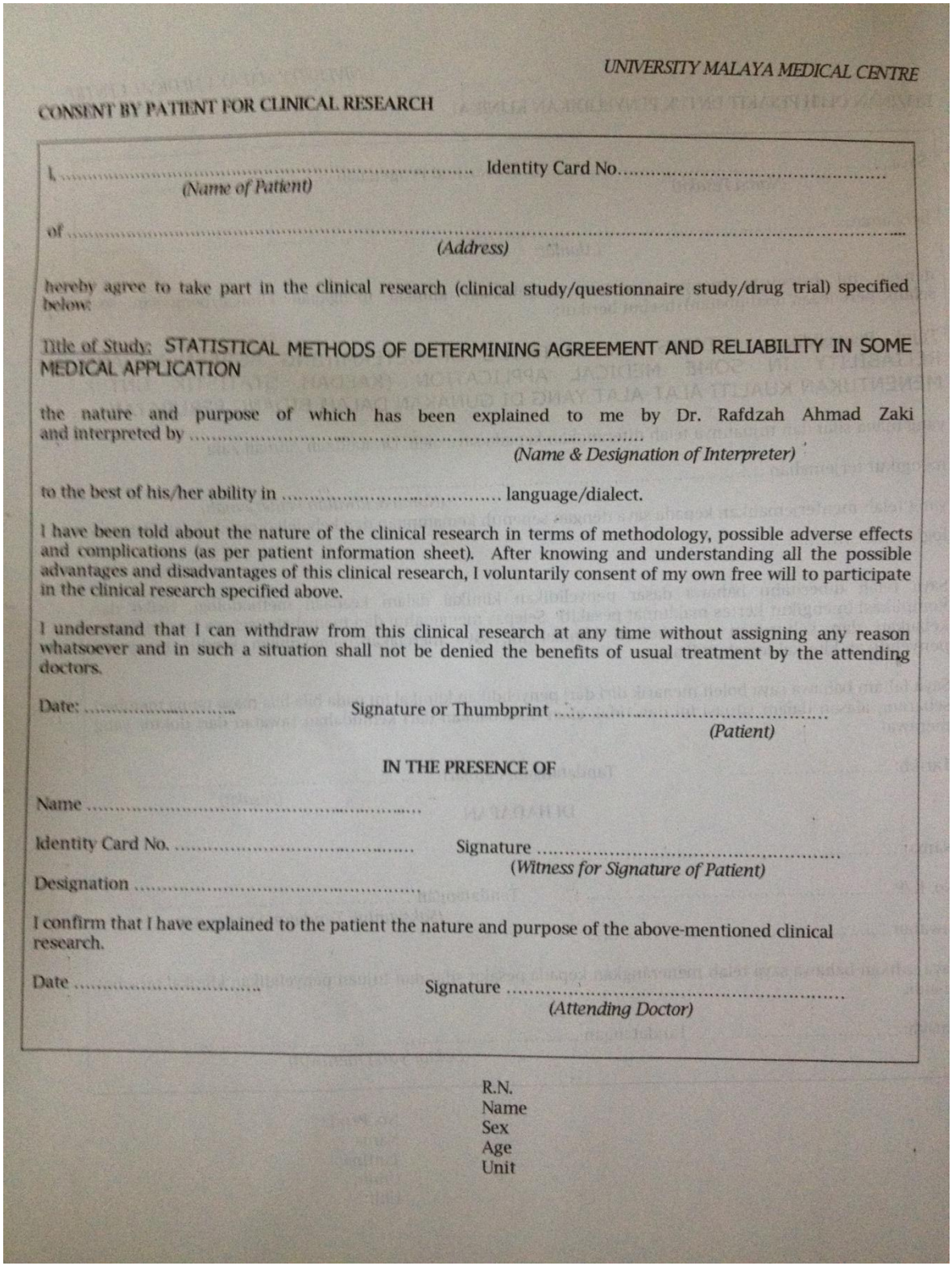
1	Tajuk	:	Statistical Methods of Detemining Agreement and Reliability in Some Medical Application
2	No. Akaun	:	PS162/2009B
3	Password	:	bgcASXue
4	Tempoh	:	18 Ogos 2009 – 17 Ogos 2010

PECAHAN		**RM**
Alat Khusus & Aksesori		6060
Bekalan		3350
Persidangan	Perjalanan	0
Kerja Lapangan	Perjalanan	500
Elaun & Gaji pekerja		6000
JUMLAH		**15910**

Sila sahkan penerimaan tawaran dengan mengembalikan borang penerimaan tawaran IPPP/UPGP/Geran (RU/PPP)/2009B terlampir **selewat-lewatnya 28 Ogos 2009 (Jumaat).** Akaun anda diaktifkan setelah pihak kami menerima jawapan penerimaan tawaran tersebut. **Jika tiada sebarang maklum balas diterima sehingga tarikh tersebut, tawaran ini akan terbatal dengan sendirinya**.

Unit Pengurusan Geran Penyelidikan
Institut Pengurusan dan Pemantauan Penyelidikan, A205 Bangunan IPS, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
Tel: (603) 7967 4522 / 4647 / 4652 / 4653 / 4654 / 4675 / 4521 / 6952 • **Faks**: (603) 7967 4648
Emel: ketua_upd_ippp@um.edu.my • http://www.ippp.um.edu.my

APPENDIX D: Consent form



UNIVERSITY MALAYA MEDICAL CENTRE

CONSENT BY PATIENT FOR CLINICAL RESEARCH

I, ... Identity Card No.....
(Name of Patient)

of .....
(Address)

hereby agree to take part in the clinical research (clinical study/questionnaire study/drug trial) specified below:

Title of Study: STATISTICAL METHODS OF DETERMINING AGREEMENT AND RELIABILITY IN SOME MEDICAL APPLICATION

the nature and purpose of which has been explained to me by Dr. Rafdzah Ahmad Zaki and interpreted by .....
(Name & Designation of Interpreter)

to the best of his/her ability in ..... language/dialect.

I have been told about the nature of the clinical research in terms of methodology, possible adverse effects and complications (as per patient information sheet). After knowing and understanding all the possible advantages and disadvantages of this clinical research, I voluntarily consent of my own free will to participate in the clinical research specified above.

I understand that I can withdraw from this clinical research at any time without assigning any reason whatsoever and in such a situation shall not be denied the benefits of usual treatment by the attending doctors.

Date: ..... Signature or Thumbprint .....
(Patient)

IN THE PRESENCE OF

Name .....

Identity Card No. ..... Signature .....
(Witness for Signature of Patient)

Designation .....

I confirm that I have explained to the patient the nature and purpose of the above-mentioned clinical research.

Date ..... Signature .....
(Attending Doctor)

R.N.
Name
Sex
Age
Unit

UNIVERSITY MALAYA MEDICAL CENTRE

KEIZINAN OLEH PESAKIT UNTUK PENYELIDIKAN KLINIKAL

Saya,..... No. Kad Pengenalan .....
(Nama Pesakit)

beralamat.....
(Alamat)

dengan ini bersetuju menyertai dalam penyelidikan klinikal (pengajian klinikal/pengajian soal-selidik/percubaan ubat-ubatan) disebut berikut:

Tajuk Penyelidikan: STATISTICAL METHODS OF DETERMINING AGREEMENT AND RELIABILITY IN SOME MEDICAL APPLICATION (KAEDAH STATISTIK UNTUK MENENTUKAN KUALITI ALAT-ALAT YANG DI GUNAKAN DALAM BIDANG PERUBATAN).

yang mana sifat dan tujuannya telah diterangkan kepada saya oleh Dr. Rafdzah Ahmad Zaki

mengikut terjemahan .....
(Nama & Jawatan Penterjemah)

yang telah menterjemahkan kepada saya dengan sepenuh kemampuan dan kebolehannya di dalam Bahasa / loghat.....

Saya telah diberitahu bahawa dasar penyelidikan klinikal dalam keadaan methodologi, risiko dan komplikasi (mengikut kertas maklumat pesakit). Selepas mengetahui dan memahami semua kemungkinan kebaikan dan keburukan penyelidikan klinikal ini, saya merelakan/mengizinkan sendiri menyertai penyelidikan klinikal tersebut di atas.

Saya faham bahawa saya boleh menarik diri dari penyelidikan klinikal ini pada bila-bila masa tanpa memberi sebarang alasan dalam situasi ini dan tidak akan dikecualikan dari kemudahan rawatan dari doktor yang merawat.

Tarikh: ..... Tandatangan/Cap Jari .....
(Pesakit)

DI HADAPAN

Nama .....

No. K/P..... Tandatangan .....
(Saksi untuk Tandatangan Pesakit)

Jawatan .....

Saya sahkan bahawa saya telah menerangkan kepada pesakit sifat dan tujuan penyelidikan klinikal tersebut di atas.

Tarikh: ..... Tandatangan .....
(Doktor yang merawat)

No. Pend.:
Nama:
Jantina:
Umur:
Unit:

APPENDIX E: Patient information sheet

UNIVERSITY OF MALAYA
The Leader in Research & Innovation

MAKLUMAT TENTANG PENYELIDIKAN

Sila baca maklumat di bawah. Sekiranya anda mempunyai soalan, sila ajukan kepada doktor yang menjalankan kajian ini.

Tajuk kajian: STATISTICAL METHODS OF DETERMINING AGREEMENT AND RELIABILITY IN SOME MEDICAL APPLICATION (KAEDAH STATISTIK UNTUK MENENTUKAN KUALITI ALAT-ALAT YANG DI GUNAKAN DALAM BIDANG PERUBATAN)

Pendahuluan:

- Ketepatan alat-alat dalam bidang perubatan adalah penting. Contohnya alat untuk menentukan tekanan darah dan paras gula dalam darah.
- Keputusan yang tidak tepat dari alat yang kurang berkualiti akan membahayakan pesakit.
- Beberapa kaedah statistik telah digunakan untuk menentukan kualiti alat-alat perubatan.
- Walaubagaimanapun tiada panduan mengenai kaedah terbaik yang patut digunakan.

Apakah tujuan penyelidikan ini?

- Tujuan penyelidikan ini ialah mencadangkan kaedah statistik baru untuk menentukan kualiti alat-alat perubatan, dan membuat perbandingan dengan kaedah yang lain.
- Penyelidikan ini juga bertujuan untuk memberi panduan mengenai kaedah terbaik untuk menentukan kualiti alat-alat perubatan.

Apakah prosedur yang akan anda lalui?

- Pengambilan tekanan darah sebanyak lima kali dengan jarak 15 saat setiap satu. Dua bacaan pertama akan diakukan secara manual. Bacaan yang seterusnya menggunakan alat automatik.
- Beberapa titik darah akan diambil dengan cucukan jarum kecil untuk mengukur paras gula di dalam darah.

Siapa yang tidak sesuai menyertai penyelidikan ini?

- Kanak-kanak dibawah 12 tahun.
- Saiz tangan yang tidak bersesuaian dengan alat tekanan darah yang digunakan.

Apakah faedah penyelidikan ini?

(a) Kepada anda sebagai peserta:

- Nasihat kesihatan secara am
- Pemeriksaan kesihatan tekanan darah dan paras gula dalam darah
- Sekiranya terdapat masalah dengan keputusan anda, kami akan memberikan nasihat.

(b) Kepada penyelidik:

- Memberikan data yang penting untuk perbandingan kaedah statistik untuk menentukan kualiti alat-alat perubatan
- Membantu kepada penyelidik menyiapkan penyelidikan

Apakah masalah yang mungkin anda hadapi?

- Sedikit tidak keselesaan
  - pengambilan tekanan darah yang berulang kali.
  - Pengambilan sampel darah

Bolehkan saya tidak menyertai penyelidikan ini?

- Anda bebas membuat keputusan samada untuk menyertai penyelidikan ini atau tidak.
- Anda juga boleh untuk mengundurkan diri pada bila-bila masa sahaja.

Siapakah patut saya hubungi jika saya ada sebarang pertanyaan mengenai penyelidikan ini?

Nama doktor: Dr Raﬁdzah Ahmad Zaki (MBChB, MPH)
Alamat:
Jabatan Perubatan Pencegahan & Kemasyarakatan,
Fakulti Perubatan, Universiti Malaya,
50603 Kuala Lumpur, MALAYSIA
Telefon: 012-5159447

Dicetak oleh: Arif Corporation Sdn. Bhd. No. 42, Jalan Pengasah, Sekyen 15/13, 40200 Shah Alam, Selangor. 03-55137242

UNIVERSITY OF MALAYA

The Leader in Research & Innovation

PATIENT INFORMATION SHEET

Please read the following information carefully, do not hesitate to discuss any questions you may have with your Doctor.

Study Title: STATISTICAL METHODS OF DETERMINING AGREEMENT AND RELIABILITY IN SOME MEDICAL APPLICATION

Introduction:

- In medicine, we are expected to have precise values of different variables. For example, accurate measurement of blood pressure, heart rate, and blood glucose level.
- Inaccurate measurement of these variables will results in inappropriate management of patient, thus will put patient's life at risk.
- Number of methods have been used to measure the quality of medical instrument.
- However, there is no clear guide or recommendation on which is the most appropriate method to use.

What is the purpose of this study?

- The purpose of this study is to propose a new statistical method to test for the quality of medical instrument, and to compare it with other methods.
- This study also aim to make recommendation on which method is the most appropriate method to be used.

What are the procedures to be followed?

- Repeated Blood Pressure measurement for five times with 15 seconds interval. First two measurement using manual method and the other three with automatic blood pressure machine.
- Very tiny prick of your finger – to obtain few drops of your blood for blood glucose measurement.

Who should not enter the study?

- Children age below 12 years old.
- Size of arm does not fit the size of cuff for Blood Pressure Measurement Tool used.

What will be benefits of the study?

(a) To you as the subject:

- General health advice.
- Health screening: Blood pressure and blood glucose level.
- If any problem with your result we will give you the appropriate advise.

(b) To the investigator:

- It will provide valuable data for statistical analysis comparison.
- It will contribute to the appropriate analysis to test the quality of medical instrument.
- It will help investigator completed this study.

What are the possible drawbacks?

- Only some discomfort on your arm due to repeated measurement of blood pressure, and very tiny prick of your finger

Can I refuse to take part in the study?

- You are free to make the decision whether you want to join the study or not.
- You are also permitted to withdraw from this study at any time during the study period.

Who should I contact if I have additional questions during the course of the study?

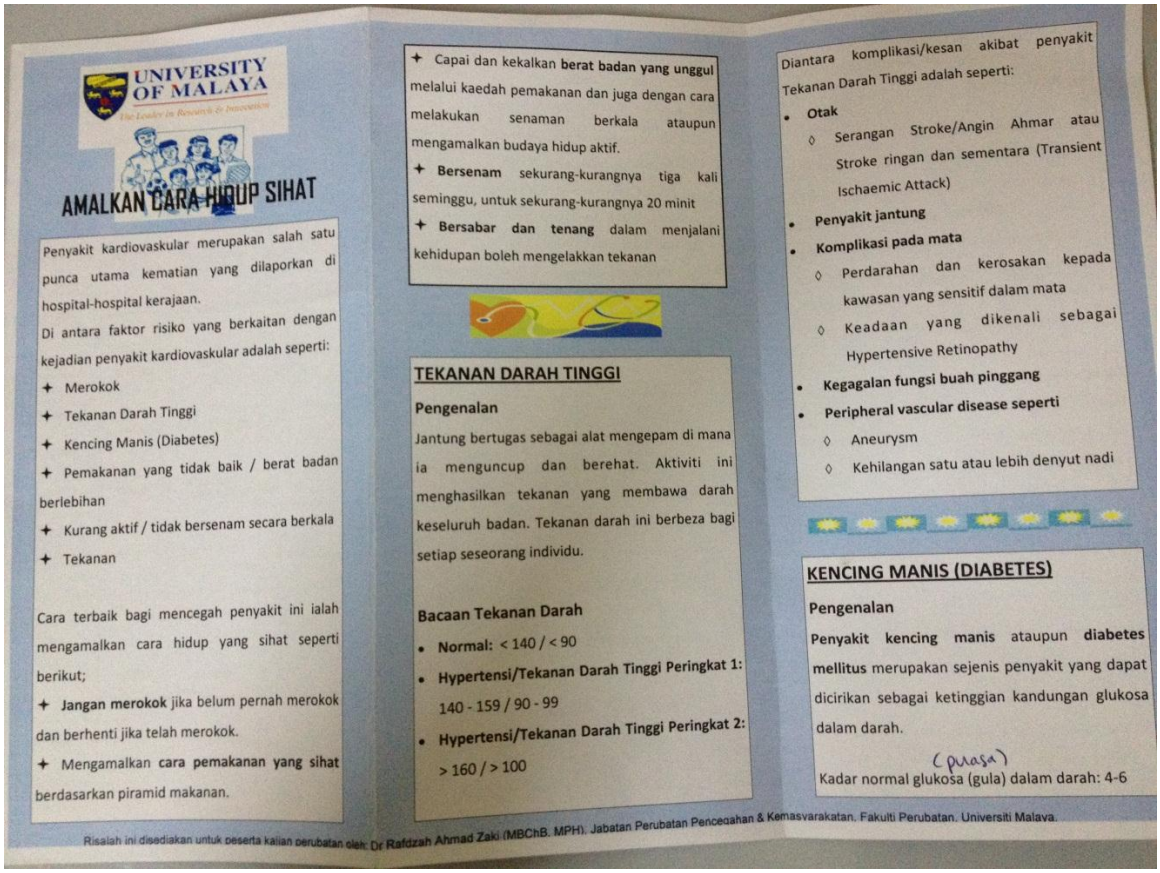
Doctor's Name:
Dr Rafdzah Ahmad Zaki (MBChB, MPH)

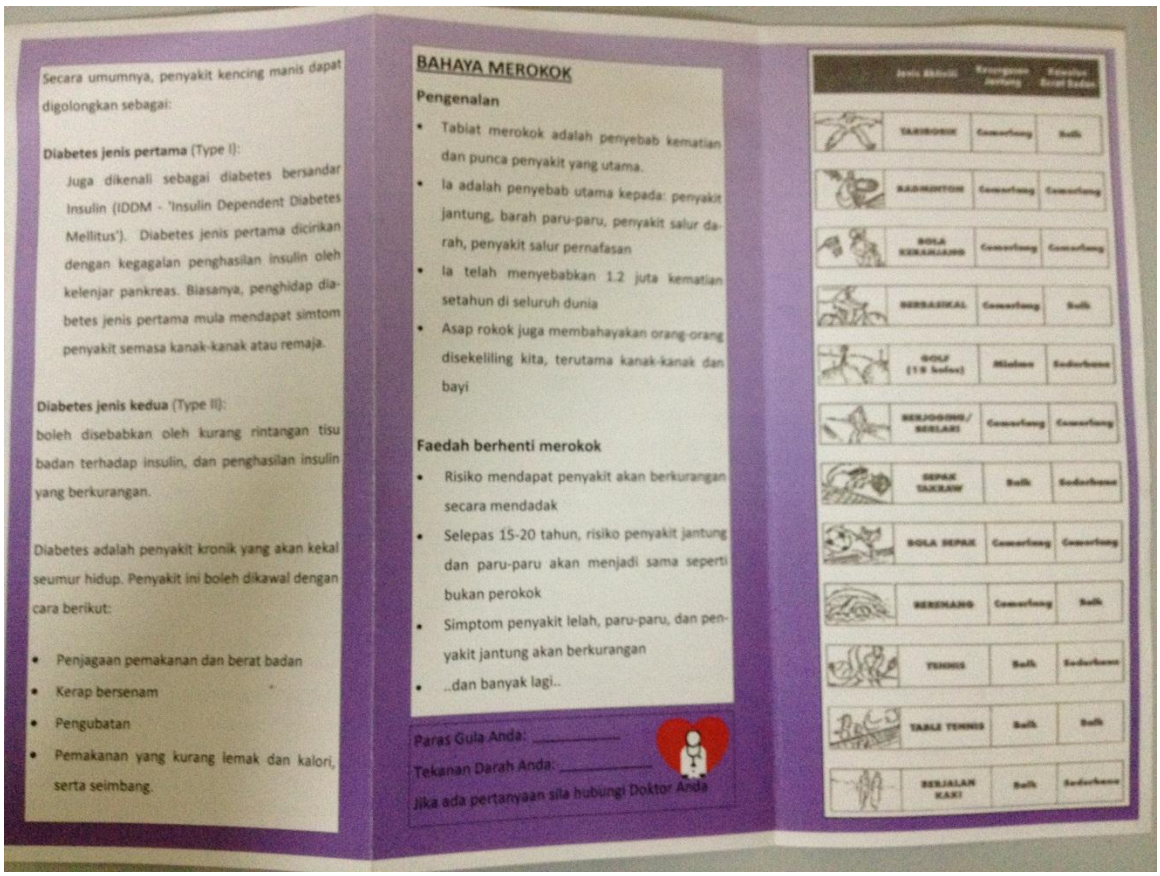
Address:
Department of Social & Preventive Medicine, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, MALAYSIA.

Tel: 012-5159447

Printed By: Arif Corporation Sdn. Bhd. No. 42, Jalan Pengasah, Seksyen 15/13, 40200 Shah Alam, Selangor 03-55137242

APPENDIX F: General health promotion leaflet





Brief description of health promotion leaflets:

- General information on risk factors for cardiovascular disease: smoking, hypertension, diabetes, poor diet, lack of exercise, and stress.
- Brief introduction on hypertension, diabetes mellitus and danger of smoking
- Examples of activities that is good for cardiovascular system and for weight reduction.

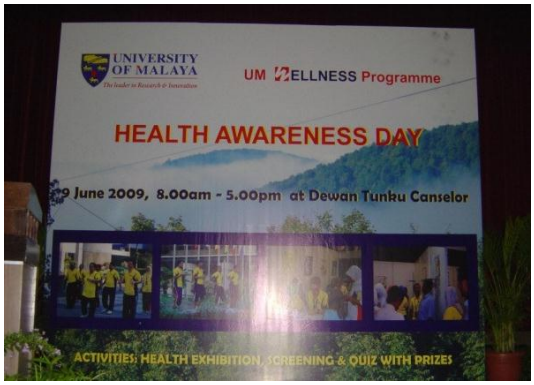
Leaflets from the Ministry of Health Malaysia:

- Stop smoking campaign
- Healthy lifestyle campaign
    - Healthy eating
    - Active
    - Not smoking
    - Good stress management
    - Do not take alcohol

APPENDIX G: Selected photos during data collection

1. UM Wellness Health Screening Programme

   a. Launching of the programme for the year 2009



   b. Blood glucose testing


   c. Blood pressure measurement


2. UM Wellness Quit Smoking Clinic

a. The Quit Smoking Clinic

b. Some of instruments used for data collection

c. Consent and measurement taken from participants

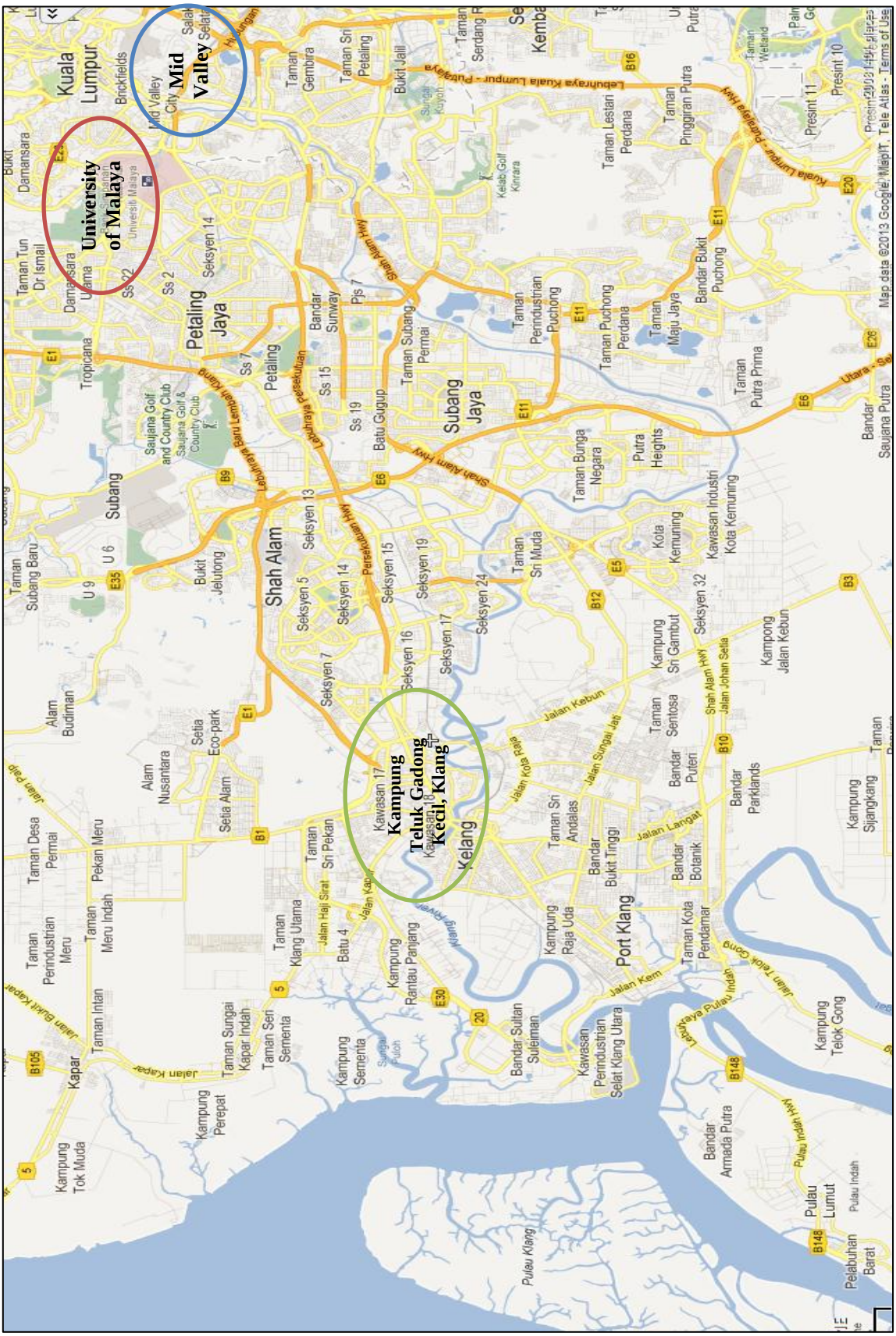
3. Community Health Screening programme, Klang

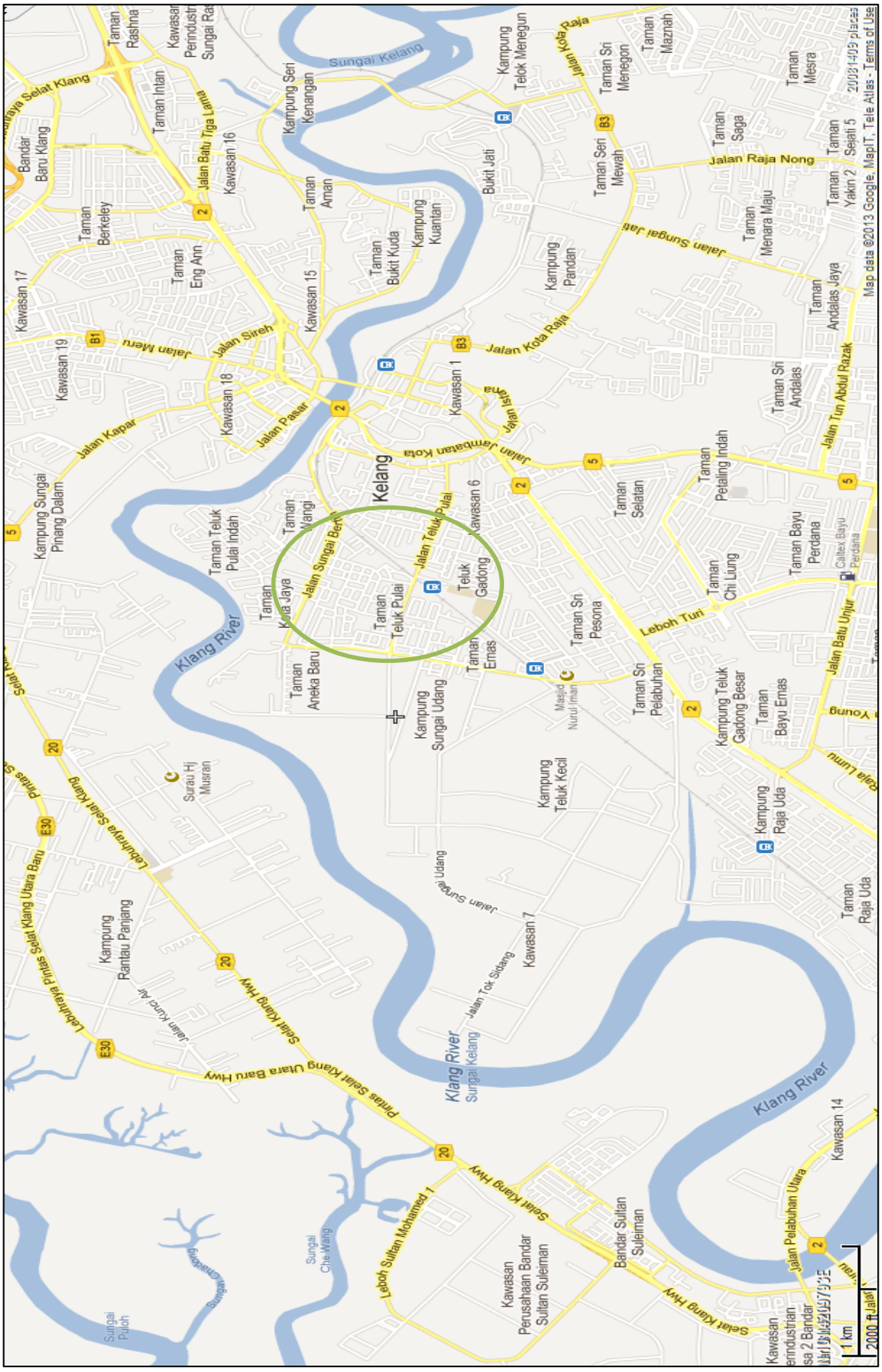
a. Pictures of community representatives and some of participants in the village

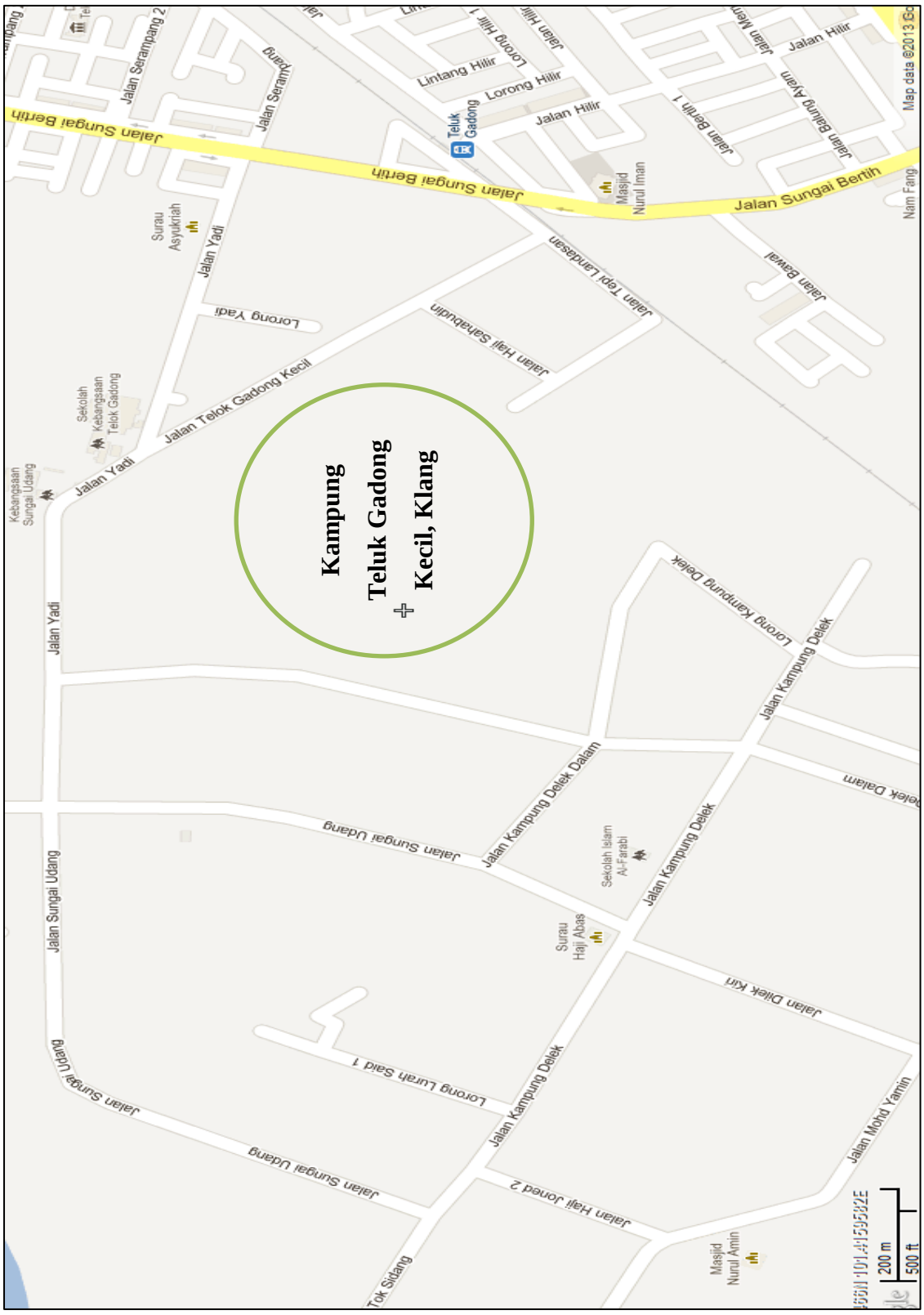
b. One of the local houses used as a centre for screening programme

c. Example of village house in the community

APPENDIX H: Map of study areas




Kelang
Klang River
Sungai Kelang
Jalan Kota Raja
Jalan Raja Nong
Jalan Kapar
Jalan Pasar
Jalan Sungai Udang
Jalan Teluk Pulai
Jalan Sungai Bera
Leboh Turi
Jalan Batu Unjur
Jalan Tun Abdul Razak
Jalan Pelabuhan Utara
Leboh Sultan Mohamed 1
Persiaran Selat Klang Utara Baru Hwy
Lebuhraya Selat Klang
Selat Klang Hwy
Kampung Sungai Udang
Kampung Teluk Kecil
Kampung Teluk Gadong Besar
Kampung Raja Uda
Kampung Rantau Panjang
Kampung Sungai Puloh
Kawasan Perusahaan Bandar Sultan Suleiman
Bandar Sultan Suleiman
Taman Teluk Pulai
Taman Teluk Pulai Indah
Taman Aneka Baru
Taman Emas
Taman Selatan
Taman Sri Pelabuhan
Taman Sri Pesona
Taman Petaling Indah
Taman Bayu Perdana
Taman Chi Liung
Taman Andalas Jaya
Taman Sri Andalas
Taman Sri Menegon
Taman Seri Mewah
Taman Saga
Taman Menara Maju
Taman Yakin 2
Taman Sejati 5
Taman Mesra
Taman Maznah
Taman Rashnia
Taman Intan
Taman Berkeley
Taman Eng Ann
Taman Aman
Taman Bukit Kuda
Taman Wangi
Kawasan 1
Kawasan 6
Kawasan 7
Kawasan 14
Kawasan 15
Kawasan 16
Kawasan 17
Kawasan 18
Kawasan 19
Kampung Seri Kenangan
Kampung Kuantan
Kampung Pandan
Kampung Telok Menegun
Kampung Sungai Pinang Dalam
Bukit Jati
Masjid Nurul Iman
Surau Hj Musran
Caltex Bayu Perdana
Map data ©2013 Google, MapIT, Tele Atlas - Terms of Use
1 km
2000 ft


Kampung
Teluk Gadong
&
Kecil, Klang
Jalan Sungai Bertih
Jalan Hilir
Lintang Hilir
Lorong Hilir
Teluk Gadong
Jalan Telok Gadong Kecil
Jalan Yadi
Jalan Kampung Delek
Jalan Kampung Delek Dalam
Jalan Sungai Udang
Jalan Mohd Yamin
Jalan Hafi Jored 2
Lorong Lurah Said 1
Tok Sidang
Masjid Nurul Iman
Masjid Nurul Amin
Surau Asyukriah
Surau Haji Abas
Sekolah Kebangsaan Telok Gadong
Sekolah Islam Al-Farabi
Jalan Tepi Lembaran
Jalan Hafi Sahabudin
Jalan Bawal
Jalan Dilek Kiri
200 m
500 ft

APPENDIX I: PRISMA Checklist 1

Section/topic	#	Checklist item	Reported on page #
TITLE: **A systematic review of statistical method used to test for agreement of medical instruments measuring continuous variable in method comparison study.**			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	25
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	81
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	25
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS). - Descriptive review only	25
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	N/A
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	26
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	26
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	27
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	28
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	29
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	26
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	29

Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means). - Percentage of statistical method used	29-32
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	N/A

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies). - Not combining results	N/A
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	28
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	30-32
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	30-32
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	N/A
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see item 15).	N/A
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see item 16]).	N/A
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	61-69, 81
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	261
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	81-82
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	yes

APPENDIX J: PRISMA Checklist 2

Section/topic	#	Checklist item	Reported on page #
TITLE : A systematic review of statistical methods used to test for reliability of medical instruments measuring continuous variables in method comparison studies.			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	50
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	81
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	50
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS). - Descriptive review only	50
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	N/A
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	50
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	50
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	51
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	51-53
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	52
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	51-52
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	52-55
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	54

		- Percentage of statistical method used	
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	N/A

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies). - Not combining results	N/A
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	51-53
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	54
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	N/A
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	54-55
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see item 15).	N/A
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see item 16]).	N/A
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	61-69,81
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	261
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	82
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	yes

APPENDIX K: Cohen's Table

TABLE 9.3.1

POWER AS A FUNCTION OF L AND u AT a = .01

	L											
u	2.00	4.00	6.00	8.00	10.00	12.00	14.00	16.00	18.00	20.00	25.00	30.00
1	12	28	45	60	72	81	88	92	95	97	99	*
2	08	20	35	49	61	72	80	87	91	94	98	99
3	07	16	29	42	54	65	74	82	87	91	97	99
4	06	14	25	37	49	60	69	77	84	89	96	98
5	05	12	22	33	44	55	65	74	80	86	94	98
6	05	11	19	30	41	51	61	70	77	83	93	97
7	04	10	18	27	37	48	58	67	74	81	91	96
8	04	09	16	25	35	45	55	64	72	78	90	96
9	04	08	15	23	33	42	52	61	69	76	88	95
10	03	08	14	22	31	40	49	58	66	74	87	94
11	03	07	13	20	29	38	47	56	64	71	85	93
12	03	07	12	19	27	36	45	54	62	69	83	92
13	03	06	12	18	26	34	43	52	60	67	82	91
14	03	06	11	17	25	33	41	50	58	65	80	90
15	03	06	10	16	23	31	40	48	56	64	79	89
16	03	06	10	16	22	30	38	46	54	62	77	88
20	02	05	08	13	19	26	33	41	48	56	72	84
24	02	04	07	12	17	22	29	36	43	51	67	80
28	02	04	07	10	15	20	26	32	39	46	62	76
32	02	04	06	09	13	18	22	29	32	42	58	72
40	02	03	05	08	11	15	20	25	30	36	51	65
50	02	03	05	07	09	13	16	21	25	31	44	58
60	02	03	04	06	08	11	14	18	22	26	39	52
80	02	02	03	05	06	09	11	14	17	21	31	43
100	01	02	03	04	06	07	09	11	14	17	26	36

* Power greater than .995.

TABLE 9.3.2

POWER AS A FUNCTION OF L AND u AT a = .05

	L											
u	2.00	4.00	6.00	8.00	10.00	12.00	14.00	16.00	18.00	20.00	25.00	30.00
1	29	52	69	81	89	93	96	98	99	99	*	*
2	23	42	58	72	82	88	93	96	97	99	*	*
3	19	36	52	65	76	84	90	93	96	98	99	*
4	17	32	47	60	72	80	87	91	94	96	99	*
5	16	29	43	56	68	77	84	89	93	95	98	*
6	15	27	40	53	64	74	81	87	91	94	98	99
7	14	25	38	50	61	71	79	85	89	93	97	99
8	13	24	36	48	59	68	77	83	88	92	97	99
9	13	23	34	45	56	66	74	81	86	90	96	99
10	12	21	32	43	54	64	72	79	85	89	96	98
11	12	21	31	42	52	64	70	78	83	88	95	98
12	11	20	30	40	50	60	69	76	82	87	94	98
13	11	19	29	39	49	58	67	74	80	85	93	97
14	11	18	28	37	47	57	65	73	79	84	93	97
15	11	18	27	36	46	55	64	71	78	83	92	97
16	10	17	26	35	45	54	62	70	76	82	91	96
20	10	16	23	31	40	49	57	65	72	78	88	94
24	09	15	21	29	37	45	53	60	67	74	85	92
28	09	14	20	27	34	42	49	57	64	70	82	91
32	08	13	18	25	32	39	46	53	60	67	80	88
40	08	12	17	22	28	40	41	48	55	61	74	84
50	08	11	15	20	25	31	37	43	49	55	69	80
60	07	10	14	18	23	28	33	39	45	50	64	75
80	07	09	12	16	20	24	28	33	38	43	56	67
100	07	09	11	14	18	21	25	29	34	38	50	61

* Power greater than .995.

APPENDIX L: Matlab Syntax (general syntax)

1. Bland-Altman analysis

```
samplesize = 500;

ULA = zeros(2,1);
LLA = zeros (2,1);
Bias = zeros (2,1);

ULActr = 1;
LLActr = 1;
Biasctr = 1;

for n =10:samplesize;

    for nctr = 1:n
        ctr(nctr) = round(1+(300-1)*rand);
    end

    for nctr = 1:n
        nsample(nctr,:) = Population2(ctr(nctr),:);
    end

    DS = sum((nsample(:,1)-(nsample(:,2))));
    mnDS= DS/n;
    ds  =  sum(((nsample(:,1))-(nsample(:,2))-mnDS).*((nsample(:,1))-
(nsample(:,2))-mnDS));
    sd = (ds/(n-1))^0.5;


    ULA(ULActr)= mnDS+1.96*sd;
    LLA(LLActr)= mnDS-1.96*sd;
    Bias(Biasctr) = mnDS;

    ULActr = ULActr + 1;
    LLActr = LLActr + 1;
    Biasctr = Biasctr + 1;

end
```

1. ICC_A

```
samplesize = 500;

ICC = zeros (2,1);
ICCctr = 1;


for n=10:samplesize;

    for nctr = 1:n
        ctr(nctr) = round(1+(300-1)*rand);
    end

    for nctr = 1:n
```

```
        nsample(nctr,:) = Population2(ctr(nctr),:);
    end

   p = anova1 (Population2);
   SSR = table{3,2};
   SSE = table{4,2};
   SSC = table{2,2};
   SSW = SSE + SSC;

   MSR = SSR / (n-1);
   MSE = SSE / ((n-1)*(k-1));
   MSC = SSC / (k-1);
   MSW = SSW / (n*(k-1));

   ICC = (MSR - MSE) / (MSR + (k-1)*MSE + k*(MSC-MSE)/n);
   ICCctr = ICC + 1;

end
```

2. Slope and Intercept

```
samplesize = 500;

slope = zeros(2,1);
intercept = zeros(2,1);
r = zeros (2,1);

interceptctr = 1;
slopectr = 1;
rctr = 1;

for n=10:samplesize;

    for nctr = 1:n
        ctr(nctr) = round(1+(300-1)*rand);
    end

    for nctr = 1:n
        nsample(nctr,:) = Population2(ctr(nctr),:);
    end

    XY = sum(nsample(:,1).*nsample(:,2));
    X2 = sum(nsample(:,1).*nsample(:,1));
    Y2 = sum(nsample(:,2).*nsample(:,2));
    X = sum(nsample(:,1));
    Y = sum(nsample(:,2));
    A = XY-((X*Y)/n);
    B = X2-(X*X/n);
    C = Y2-(Y*Y/n);


    slope(slopectr) = ((n*XY)-(X*Y))/((n*X2)-X^2);
    intercept(interceptctr)=((Y)-(((n*XY)-(X*Y))/((n*X2)-
X^2))*(X))/n ;
    r(rctr)= A/((B*C)^0.5);

    interceptctr = interceptctr + 1;
    slopectr = slopectr + 1;
    rctr = rctr + 1;
end
```

3. Agreement model (general formula)

SampleSize2 = standard value

SlopeX = slope.*SampleSize2

Predicted = SlopeX+intercept

Error = Predicted-SampleSize2

5. Error simulation(general formula): Section 4.3.3

Blood glucose & weight		*Systolic BP*	
Range of simulated error	General Matlab syntax	Range of simulated error	General Matlab syntax
Error 0 to 0.1 = 0.1	y = 0+0.1*rand(300)	Error 0 to 2.0 = 2	y= 0+2*rand(300,1)
Error 0 to 0.2 =0.2	y=0+0.2*rand(300,1)	Error 0 to 4.0 = 4	Y=0+4*rand(300,1)
Error 0 to 0.3 = 0.3	y=0+0.3*rand(300,1)	Error 0 to 6.0 = 6	y=0+6*rand(300,1)
Error 0 to 0.4 = 0.4	y=0+0.4*rand(300,1)	Error 0 to 8.0 = 8	y=0+8*rand(300,1)
Error 0 to 0.5 = 0.5	y=0+0.5*rand(300,1)	Error 0 to 10 = 10	y=0+10*rand(300,1)
Error 0 to 0.6 = 0.6	y=0+0.6*rand(300,1)	Error 0 to 12 = 12	y=0+12*rand(300,1)
Error 0 to 0.7 = 0.7	y=0+0.7*rand(300,1)	Error 0 to 14 = 14	y=0+14*rand(300,1)
Error 0 to 0.8 = 0.8	y=0+0.8*rand(300,1)	Error 0 to 16 = 16	y=0+16*rand(300,1)
Error 0 to 0.9 = 0.9	y=0+0.9*rand(300,1)	Error 0 to 18 = 18	y=0+18*rand(300,1)
Error 0 to 1.0 = 1.0	y=0+1.0*rand(300,1)	Error 0 to 20 = 20	y=0+20*rand(300,1)
Error -0.1 to 0.1 = 0.2	y=-0.1+0.2*rand(300,1)	Error -2 to 2 = 4	y=-2+4*rand(300,1)
Error -0.2 to 0.2 = 0.4	y=-0.2+0.4*rand(300,1)	Error -4 to 4 = 8	y=-4+8*rand(300,1)
Error -0.3 to 0.3 = 0.6	y=-0.3+0.6*rand(300,1)	Error -6 to 6 = 12	y=-6+12*rand(300,1)
Error -0.4 to 0.4 = 0.8	y=-0.4+0.8*rand(300,1)	Error -8 to 8 = 16	y=-8+16*rand(300,1)
Error -0.5 to 0.5 = 1.0	y=-0.5+1.0*rand(300,1)	Error -10 to 10 = 20	y=-10+20*rand(300,1)
Error -0.6 to 0.6 = 1.2	y=-0.6+1.2*rand(300,1)	Error -12 to 12 = 24	y=-12+24*rand(300,1)
Error -0.7 to 0.7 = 1.4	y=-0.7+1.4*rand(300,1)	Error -14 to 14 = 28	y=-14+28*rand(300,1)
Error -0.8 to 0.8 = 1.6	y=-0.8+1.6*rand(300,1)	Error -16 to 16 = 32	y=-16+32*rand(300,1)
Error -0.9 to 0.9 = 1.8	y=-0.9+1.8*rand(300,1)	Error -18 to 18 = 36	y=-18+36*rand(300,1)
Error -1.0 to 1.0 = 2.0	y=-1.0+2.0*rand(300,1)	Error -20 to 20 = 40	y=-20+40*rand(300,1)

APPENDIX M: Proof of publication

1. Zaki, R., Bulgiba, A., Ismail, R., & Ismail, N. A. (2012). Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review. *PLoS ONE, 7: e37908.doi:10.1371/journal.pone.0037908*

OPEN ACCESS Freely available online

PLoS one

Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review

Rafdzah Zaki[1]*, Awang Bulgiba[1], Roshidi Ismail[1], Noor Azina Ismail[2]

1 Julius Centre University of Malaya, Department of Social and Preventive Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia, **2** Department of Applied Statistics, Faculty of Economics and Administration, University of Malaya, Kuala Lumpur, Malaysia

Abstract

Background: Accurate values are a must in medicine. An important parameter in determining the quality of a medical instrument is agreement with a gold standard. Various statistical methods have been used to test for agreement. Some of these methods have been shown to be inappropriate. This can result in misleading conclusions about the validity of an instrument. The Bland-Altman method is the most popular method judging by the many citations of the article proposing this method. However, the number of citations does not necessarily mean that this method has been applied in agreement research. No previous study has been conducted to look into this. This is the first systematic review to identify statistical methods used to test for agreement of medical instruments. The proportion of various statistical methods found in this review will also reflect the proportion of medical instruments that have been validated using those particular methods in current clinical practice.

Methodology/Findings: Five electronic databases were searched between 2007 and 2009 to look for agreement studies. A total of 3,260 titles were initially identified. Only 412 titles were potentially related, and finally 210 fitted the inclusion criteria. The Bland-Altman method is the most popular method with 178 (85%) studies having used this method, followed by the correlation coefficient (27%) and means comparison (18%). Some of the inappropriate methods highlighted by Altman and Bland since the 1980s are still in use.

Conclusions: This study finds that the Bland-Altman method is the most popular method used in agreement research. There are still inappropriate applications of statistical methods in some studies. It is important for a clinician or medical researcher to be aware of this issue because misleading conclusions from inappropriate analyses will jeopardize the quality of the evidence, which in turn will influence quality of care given to patients in the future.

Citation: Zaki R, Bulgiba A, Ismail R, Ismail NA (2012) Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review. PLoS ONE 7(5): e37908. doi:10.1371/journal.pone.0037908

Editor: Fabio Rapallo, University of East Piedmont, Italy

Received February 20, 2012; **Accepted** April 30, 2012; **Published** May 25, 2012

Funding: This work is part of the Meta-Analysis project (STeMM Programme) supported by the University of Malaya/Ministry of Higher Education (UM/MOHE) High Impact Research Grant (Grant number E000010-20001), and the University of Malaya student research grant (PS162/2009B). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rafdzah@hotmail.com

Introduction

Most important variables in medicine are measured in numerical forms or continuous data, such as blood pressure, glucose level and oxygen level. In any clinical situation, we are expected to have accurate readings of these variables. Numerous new techniques or tools have been developed with the aim of finding a cheaper, non-invasive, more convenient and safer method to test patients. It is important to be sure that the new tool or method of measurement is as accurate as the current or gold standard method. Therefore it is important to measure the agreement of the new method with the standard method. Agreement signifies the accuracy of that certain instrument [1].

Various statistical methods have been used to test for agreement of medical instruments with quantitative or continuous outcomes [2,3]. Which method is the best is still open to debate and almost all methods have been criticized. The old favorite for measuring agreement is the correlation coefficient (r) [4]. However, this is obviously inappropriate as correlation only measures the strength of linear association between variables. Coefficient of determination (r^2), regression coefficient, and comparing means have also been shown to be inappropriate ways of assessing agreement. This was discussed by Altman and Bland in their article [2] back in the 1980s. Their conclusions on the inappropriate methods to assess agreement have been supported by Daly and Bourke [4], and there is little argument about this in the literature.

Bland and Altman proposed a method for the analysis of agreement (Bland-Altman plot and limits of agreement) in 1983 [2] and later drew the attention of the medical profession to this area in their article [5] in the Lancet. They stated that it is very unlikely for two different methods or instrument to be exactly in agreement, or to give identical results for all individuals [5]. What is important is how close the pairs of values are [5]. This is because a very small difference in the predicted and the actual value is not

2. Zaki, R., Bulgiba, A., & Ismail, N. A. Testing the agreement of medical instruments: Overestimation of bias in the Bland-Altman analysis. Accepted by Preventive Medicine journal.

Preventive Medicine

Available online 10 January 2013

In Press, Corrected Proof — Note to users

Testing the agreement of medical instruments: Overestimation of bias in the Bland–Altman analysis

Rafdzah Zaki[a], , Awang Bulgiba[a], Noor Azina Ismail[b]

[a] Julius Centre University of Malaya, Department of Social & Preventive Medicine, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia

[b] Department of Applied Statistics, Faculty of Economics & Administration, University of Malaya, 50603, Kuala Lumpur, Malaysia

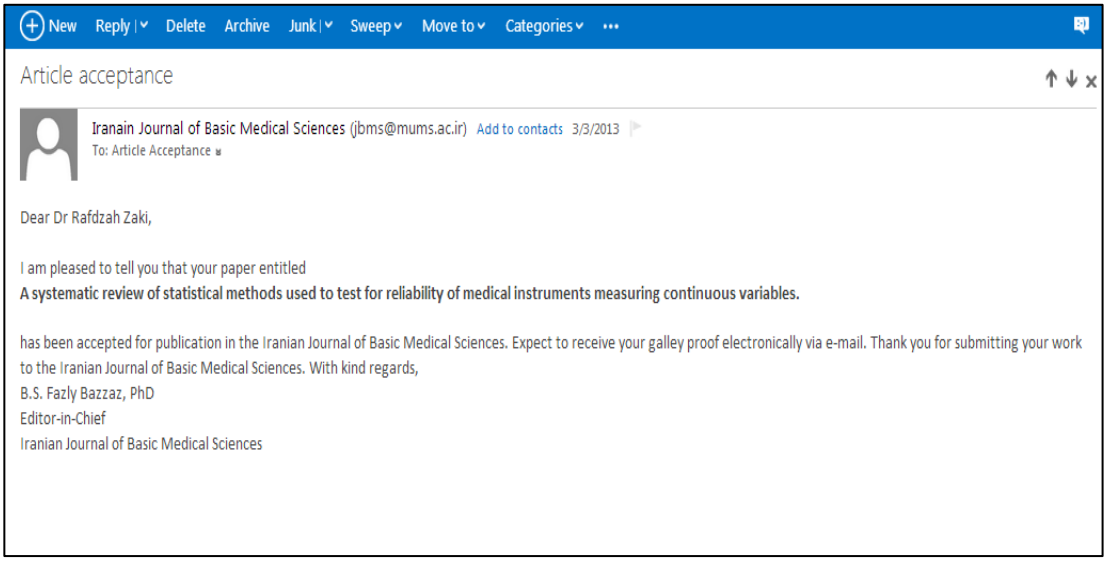
http://dx.doi.org/10.1016/j.ypmed.2013.01.003, How to Cite or Link Using DOI

Permissions & Reprints

View full text

Purchase $31.50

3. Zaki, R., Bulgiba, A., Nordin, N & Ismail, N. A. A systematic review of statistical methods used to test for reliability of medical instruments measuring continuous variables. Accepted by Iranian Journal of Basic Medical Science.



New Reply Delete Archive Junk Sweep Move to Categories

Article acceptance

Iranain Journal of Basic Medical Sciences (jbms@mums.ac.ir) Add to contacts 3/3/2013

To: Article Acceptance

Dear Dr Rafdzah Zaki,

I am pleased to tell you that your paper entitled

A systematic review of statistical methods used to test for reliability of medical instruments measuring continuous variables.

has been accepted for publication in the Iranian Journal of Basic Medical Sciences. Expect to receive your galley proof electronically via e-mail. Thank you for submitting your work to the Iranian Journal of Basic Medical Sciences. With kind regards,

B.S. Fazly Bazzaz, PhD

Editor-in-Chief

Iranian Journal of Basic Medical Sciences